Using Comparable Corpora to Augment Low Resource SMT Models

Ann Irvine

December 10, 2014

Contents

1 Introduction **7**
- 1.1 Contributions of this thesis . . . 11
- 1.2 Structure of this document . . . 11
- 1.3 Related publications . . . 13

2 Literature Review **15**
- 2.1 Statistical Machine Translation . . . 15
  - 2.1.1 Phrase-Based Machine Translation . . . 15
  - 2.1.2 Other Models of Translation . . . 16
  - 2.1.3 Low Resource Machine Translation . . . 18
    - 2.1.3.1 AVENUE . . . 18
    - 2.1.3.2 METIS-II . . . 19
    - 2.1.3.3 Other work . . . 19
  - 2.1.4 The OOV and rare word problem in SMT . . . 19
- 2.2 Expanding bilingual resources . . . 20
  - 2.2.1 Bilingual lexical induction . . . 20
    - 2.2.1.1 Contextual Similarity . . . 20
    - 2.2.1.2 Other Monolingual Similarity Metrics . . . 21
    - 2.2.1.3 Integration with SMT . . . 22
  - 2.2.2 Extracting Parallel Data from Comparable Corpora . . . 22
  - 2.2.3 Crowdsourcing Translations . . . 22
  - 2.2.4 Automatically Expanding SMT Model Coverage . . . 23
- 2.3 Domain Adaptation for Machine Translation . . . 23

3 Languages, Data, and Analysis **25**
- 3.1 Languages . . . 25
- 3.2 Parallel Corpora and Dictionaries . . . 26
  - 3.2.1 Parallel Corpora . . . 26
  - 3.2.2 Bilingual Dictionaries . . . 27
- 3.3 Monolingual and Comparable Corpora . . . 28
  - 3.3.1 Web crawls . . . 28
  - 3.3.2 Wikipedia . . . 30
- 3.4 Analysis . . . 30
  - 3.4.1 Approach . . . 30
    - 3.4.1.1 Error Taxonomy . . . 30
    - 3.4.1.2 Word Alignment Driven Evaluation . . . 30
    - 3.4.1.3 Table Enhancement for Translation Analysis . . . 31
  - 3.4.2 Experiments . . . 32
    - 3.4.2.1 WADE Analyses . . . 32
    - 3.4.2.2 TETRA Analyses . . . 34
  - 3.4.3 Word Alignment Errors . . . 35
  - 3.4.4 Analysis Conclusion . . . 35

4 Bilingual Lexicon Induction **37**
4.1 Motivating Prior Work . . . 37
4.2 Using Monolingual Data to Predict Translations . . . 38
4.2.1 Monolingual Signals of Translation Equivalence . . . 38
4.2.1.1 Contextual Similarity . . . 38
4.2.1.2 Temporal Similarity . . . 38
4.2.1.3 Orthographic Similarity . . . 39
4.2.1.4 Topic Similarity . . . 41
4.2.1.5 Frequency Similarity . . . 43
4.2.1.6 Burstiness Similarity . . . 43
4.2.1.7 Additional Signals . . . 44
4.2.2 Orthogonality of Signals . . . 44
4.2.3 Individual Monolingual Signals . . . 47
4.2.4 Learning to combine orthogonal monolingual signals . . . 48
4.3 Experiments . . . 51
4.3.1 Comparison with Unsupervised Baseline . . . 51
4.3.2 Analysis by Word Frequency . . . 52
4.3.3 Analysis by Word Burstiness . . . 52
4.3.4 Analysis by Amount of Monolingual Data . . . 54
4.4 Learning Curve Analyses . . . 56
4.4.1 Translated Word Pairs . . . 56
4.4.2 Monolingual Data . . . 56
4.5 Learning Models Across Languages . . . 59
4.6 Comparison with Prior Work . . . 59
4.7 Conclusions . . . 61

5 Monolingual Phrase Table Scoring **63**
5.1 Phrase Table Scoring . . . 63
5.1.1 Phrasal Features . . . 63
5.1.2 Lexical Features . . . 64
5.2 Experiments with An Existing Phrase Table . . . 64
5.2.1 Results . . . 66
5.2.1.1 Ablation Experiments . . . 66
5.2.1.2 Combining Bilingually and Monolingually Estimated Features . . . 66
5.3 Phrase Table Scoring Conclusions . . . 67

6 End-to-End SMT with Zero or Small Parallel Texts **69**
6.1 Improving Coverage . . . 69
6.2 Improving Accuracy . . . 71
6.3 Zero Parallel Data Setting . . . 71
6.4 Small Parallel Corpora Setting . . . 72
6.4.1 Data . . . 72
6.4.2 Experimental setup . . . 72
6.4.3 Results . . . 75
6.4.3.1 Bilingual Lexicon Induction . . . 75
6.4.3.2 Improving Coverage and Accuracy in End-to-End SMT . . . 76
6.4.3.3 Translations of Low Frequency Words . . . 76
6.4.3.4 Appending Top-K Translations . . . 77
6.4.3.5 Learning Curves over Parallel Data . . . 77
6.4.3.6 Learning Curves over Comparable Corpora . . . 78
6.4.4 Post-Augmentation WADE Analysis . . . 78
6.4.4.1 WADE with Multiple References . . . 78
6.4.4.2 Analysis . . . 82
6.4.4.3 WADE Analysis Conclusions . . . 84

6.5 End-to-End SMT with Zero or Small Parallel Texts Conclusion . . . 84

7 Phrase Translation Mining 87
7.1 Option 1: Fast Phrase Pair Filtering . . . 89
7.2 Option 2: Composing Phrase Translations . . . 90
7.2.1 Motivating Experiments . . . 91
7.2.2 Phrase Composition Algorithm . . . 95
7.2.3 Pruning Phrase Pairs Using Scores Derived from Comparable Corpora . . . 95
7.3 End-to-end SMT with Induced Phrase Translations . . . 98
7.3.1 Experimental Setup . . . 98
7.3.2 Unigram Translations . . . 98
7.3.3 Composing and Pruning Phrase Translations . . . 98
7.3.4 MT Experimental Setup . . . 99
7.3.5 Results . . . 99
7.3.5.1 Unigram Translations . . . 99
7.3.5.2 Composed Phrase Pairs . . . 99
7.3.5.3 End-to-End Translation . . . 101
7.3.6 Discussion . . . 101

8 From Low Resource MT to Domain Adapted MT 105
8.1 Domain Adaptation Data . . . 105
8.2 WADE and TETRA Analyses . . . 106
8.3 New-Domain Comparable Corpora . . . 108
8.4 Using Comparable Corpora to Score Phrase Tables for Domain Adaptation . . . 109
8.5 Using Comparable Corpora to Translate Unseen Words for Domain Adaptation . . . 111
8.5.1 Bilingual Lexicon Induction Model . . . 111
8.5.2 Evaluation of Induced Translations . . . 112
8.5.3 Integrating Translations into End-to-End SMT . . . 113
8.5.4 Conclusion . . . 114

9 Conclusion 115

10 Language Set 119

11 Data Resources 123

12 WADE Analysis: Comparison of the use of Automatic and Manual Word Alignments 131

13 Bilingual Lexicon Induction 133
13.1 Contextual Vector Projection Dictionaries . . . 133
13.2 Comparison of Temporal Signatures . . . 133

14 Zero-Parallel Data Translations 139

15 Fast Phrase Pair Filtering 143
15.1 Exploratory Experiments in Learning Effective, Efficient Filters . . . 143
15.2 Scaling Up . . . 147

Chapter 1

Introduction

The objective of this thesis is to directly incorporate comparable corpora into the estimation of end-to-end statistical machine translation (SMT) models. Typically, SMT models are estimated from *parallel* corpora, or pairs of translated sentences. In contrast to parallel corpora, comparable corpora are pairs of monolingual corpora that have some cross-lingual similarities, for example topic or publication date, but that do not necessarily contain any direct translations. Comparable corpora are more readily available in large quantities than parallel corpora, which require significant human effort to compile. We use comparable corpora to estimate machine translation model parameters and show that doing so improves performance in settings where a limited amount of parallel data is available for training. Specifically, we use comparable corpora in the following ways:

- We induce high quality translations for words and short multiword phrases using features estimated over comparable corpora and a small number of example translations.
- We estimate a variety of phrase table feature functions over phrase pairs both extracted from parallel data and induced from comparable corpora.
- We show that the induced translations and new feature functions significantly improve the quality of machine translations for both low resource language pairs and text domains.

In the last twenty years, statistical machine translation has improved dramatically as algorithms for using parallel corpora to build translation models have improved. However, for many potential machine translation use cases, the parallel corpora necessary to train high quality statistical models do not exist. Even if we restrict ourselves to those markets with many potential users, focusing only on languages with tens of millions of speakers, there are thousands of possible language pairs. At best, we have substantial quantities of parallel corpora in a few hundred of these. Furthermore, in most of those cases, the parallel corpora consist of text from the government domain. For the vast majority of languages and domains, zero or very little parallel data exists (Lopez and Post, 2013). This thesis focuses on these critical low resource machine translation settings, where insufficient parallel corpora are available for training statistical models.

This thesis addresses two slightly different low resource settings. The first is typically referred to simply as 'low resource machine translation' and is characterized by the amount of parallel data available for training SMT models for a given *pair of languages*. In Chapter 3 we present the amount of parallel data publicly available for over one hundred languages paired with English. For Chinese and French, over one *billion* words of parallel text are available for training SMT models. Additionally, for many languages, including, for example, Italian, Serbian, Japanese, and Ukrainian, over one million words of parallel text are available. In this thesis, we generally define low resource language pairs as those for which we have access to fewer than one million words of parallel data. Low resource language pairs include, for example, Malayalam, Somali, Kazakh, and Afrikaans paired with English. Because doing evaluation on truly low resource language pairs can be difficult, in some experiments we simulate low resource conditions by training SMT models on samples of larger parallel training sets.

Figure 1.1 demonstrates the effect of using varying amounts of parallel training data to estimate a statistical machine translation model. To show the effect, we sample varying amounts of training data from the very large French-English Hansard parliamentary proceedings parallel corpus.[1] We train SMT models on each set of training data and measure

[1] This corpus consists of manual transcriptions and translations form the Canadian parliament. It is described in more detail in Chapter 8.

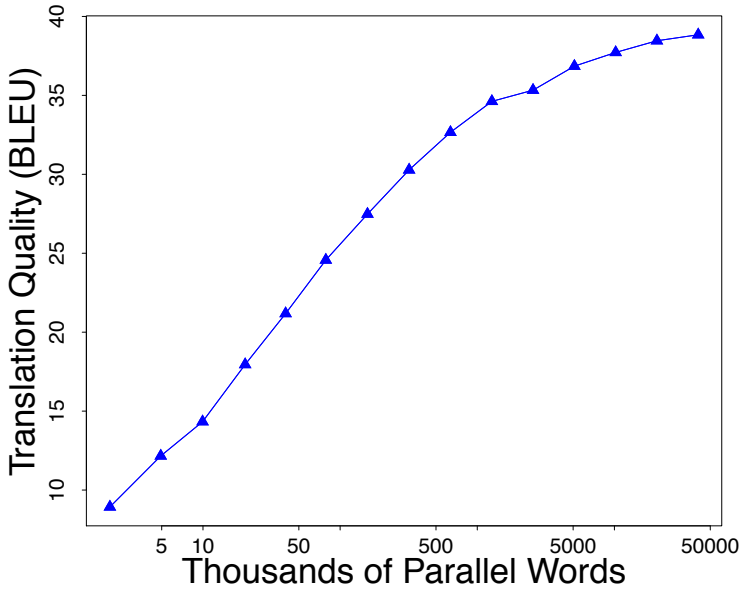


Figure 1.1: Illustration of the effect of training data size (x-axis) on translation quality (y-axis). We estimate French to English translation models using varying amounts of the very large French-English Hansard parliamentary proceedings parallel dataset. Then, we measure the performance of each model by translating a held-out French test set, also taken from the French-English Hansard parliamentary proceedings, into English and computing the BLEU score, a metric for automatically estimating translation quality.

the performance of each by translating a test set of French sentences, also taken from the Hansard corpus, and then comparing the English machine translations with reference English translations. As the amount of parallel data increases (along the x-axis), the quality of the machine translations increases. The effect is nearly linear with the log of the amount of training data.

The second low resource setting that we address in this thesis involves the *domain* of the text that we wish to translate. In many real-world translation settings we have access to parallel data in some text domain, for example parliamentary proceedings, but we wish to translate text in another domain, for example from the medical domain. Training SMT models on text in some 'old-domain' and then using the models to translate text in a 'new-domain' results in low quality translations due to the mismatch.

Figure 1.2 demonstrates the domain mismatch effect. The lower (red) bars show the results of using a French-English SMT model trained on old-domain (parliamentary proceedings) data to translate text in two new-domains.[2] When we, instead, use a combination of old-domain and new-domain parallel data to train our SMT models, performance increases substantially. In the medical domain, the BLEU score increases by about 23% and in the science domain, it increases by about 28%. The performance increases are substantial despite the fact that the old-domain parallel training data is very large (about 150 million words). Such large gaps in performance mean that adapting the old-domain models to new domains of text has the potential to yield much higher quality translations in the new-domain.

In Chapter 3 we present results from our novel error analysis methodology which shed light on the types and number of errors that are made when we train SMT models to translate between pairs of low resource languages. In Chapter 8 we present a similar analysis of what goes wrong when we use an SMT model trained on text in one domain to translate text in a different domain. In both cases, we find that most errors are due to unseen source language words and phrases and unseen target language translations. We also find room for improving errors due to how different translations are weighted, or scored, in the low resource SMT models.

In this thesis, we present and release comparable corpora in 151 languages paired with English as well as, for most languages, a bilingual dictionary of single word translations. We harvested the comparable corpora from both

[2]Details about the new domain datasets are given in Chapter 8.

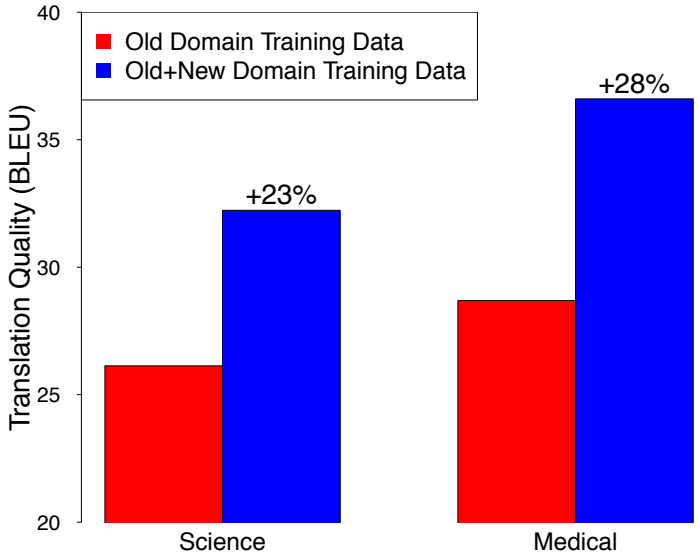


Figure 1.2: Illustration of the effect of using old-domain data to translate text in two different new-domains. We compare the performance of models trained on old-domain data only with that of models trained on both old and new domain parallel training data. Our old domain data consists of the complete French-English Hansard parliamentary proceedings dataset, and our new domain datasets are taken from the science and medical domains. When we add new-domain training data to our strong old-domain baselines, translation quality increases by 23% and 28% BLEU on the science and medical translation tasks, respectively.

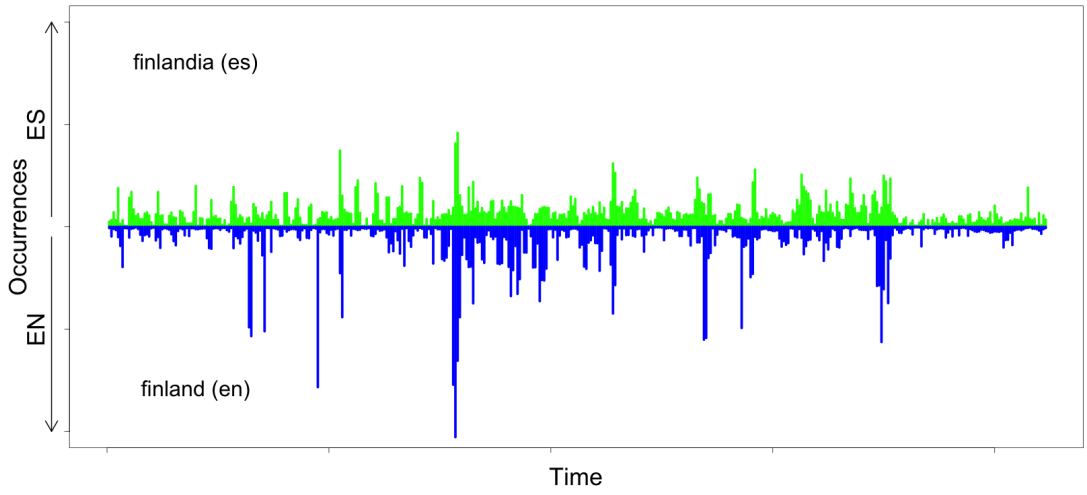


Figure 1.3: Visualization of the temporal similarity between a pair of word translations. The frequency of *finlandia* in Spanish monolingual corpora over time is similar to that of *finland* in English monolingual corpora over time.

Wikipedia and online newspapers. Our bilingual dictionaries are compiled from a variety of sources, including electronic dictionaries, scanned and digitized paper dictionaries, and crowdsourced translations gathered on Amazon's Mechanical Turk platform. We provide details on the collection and content of each data resource in Chapter 3.

Using comparable corpora, we estimate a variety of *signals of translation equivalence* and use these signals to both identify new translations and to score bilingually extracted translations. The signals estimated over comparable corpora include contextual, temporal, topic, burstiness, and frequency similarity. Temporal similarity, for example, is a measure of the similarity between temporal signatures of a source language word or phrase and a target language word or phrase. The intuition behind this signal is that news stories in different languages published around the world tend to discuss the same events as they occur. Temporal signatures are estimated using the frequencies of a given word or phrase across a set of documents, each of which is associated with some timestamp. We can align source language documents with target language documents by their timestamps and then directly compare temporal signatures. Figure 1.3 shows an example of the temporal signature of Spanish *finlandia* contrasted with that of English *finland*. The temporal signatures are estimated over our comparable corpora of time-stamped web crawls of Spanish and English online newspapers. In general, the country Finland is mentioned with similar frequency over time in both the Spanish and English corpora. This temporal similarity signal is one weak indication that the two words may be translations of one another. We estimate topic similarity in a similar way. However, instead of aligning document pairs by timestamp, we aligned them by topic. Another signal, which is perhaps more obvious in the case of *finlandia* and *finland*, is orthographic similarity. We also make use of this signal, which does not reference comparable corpora. In Chapter 4, we present each signal in detail.

Although the signals of translation equivalence that we use in this thesis have all been proposed in prior work, we present novel techniques for *combining* them to make accurate translation predictions. In particular, we define novel ways to use the signals of translation equivalence to improve translation errors due to both unseen translations and inaccurately scored translations. Prior work has induced unigram translations from comparable corpora using a variety of unsupervised learning techniques, but most of this work does not go as far as to augment end-to-end SMT models with the new translations. We present a novel *supervised* approach to inducing new translations from comparable corpora. Our approach requires only a small number of example translations for training, and we integrate the new translations into end-to-end SMT models trained in low resource conditions. We also present novel techniques for integrating scores based on comparable corpora directly into end-to-end SMT models. We show that such scores lead to improved translation quality in both of our low resource settings. Augmenting baseline SMT models with both new translations and new scores yields even bigger performance gains.

In contrast to a lot of prior work on low resource machine translation (see Section 2.1.3), we take a *language-independent approach*. This follows prior work on statistical machine translation, where the pipeline for training

models is typically the same for any language pair. Because they depend only on data resources and not expert human knowledge or hand-written rules, language independent approaches have the advantage of allowing for very quick deployment of new systems for new language pairs. This is particularly advantageous in our target setting because new low resource languages may become of interest suddenly, for example in emergency disaster situations. During the January 2010 earthquake in Haiti, human volunteers translated Creole text messages that survivors sent to English speaking relief workers. If SMT systems could be deployed very quickly in similar situations in the future, they may be able to supplement or replace such crowdsourcing efforts (Lewis *et al.*, 2011).

1.1 Contributions of this thesis

The major contributions of this thesis are as follows:

- We release 'language packs' for 151 human languages, which include bilingual dictionaries, comparable corpora of Wikipedia document pairs, comparable corpora of time-stamped news text that we harvested from the web, and, for non-roman script languages, dictionaries of name pairs, which are likely to be transliterations.
- We present a novel technique for using a small number of example word translations to learn a supervised model for bilingual lexicon induction which takes advantage of a wide variety of signals of translation equivalence that can be estimated over comparable corpora.
- We show that using comparable corpora to induce new translations and estimate new phrase table feature functions improves end-to-end statistical machine translation performance for low resource language pairs as well as domains.
- We present a novel algorithm for composing multiword phrase translations from multiple unigram translations and then use comparable corpora to prune the large space of hypothesis translations. We show that these induced phrase translations improve machine translation performance beyond that of component unigrams.

1.2 Structure of this document

Figure 1.4 gives a schematic overview of the thesis and shows how each chapter fits into a simplistic view of statistical machine translation. The thesis is structured as follows:

- Chapter 2 gives an introduction to statistical machine translation, with a focus on the phrase-based model that we use throughout the thesis. It also introduces and reviews prior work on expanding bilingual resources, including bilingual lexicon induction and crowdsourcing translations, and domain adaptation for machine translation.
- Chapter 3 introduces the set of languages and text corpora that we use throughout the thesis. This chapter also presents results from a novel error analysis technique to better understand what goes wrong in machine translation when we train models on only small amounts of parallel training data.
- Chapter 4 details our novel technique for doing supervised bilingual lexicon induction using a diverse set of similarity measures that indicate translation equivalence. This chapter includes an analysis of a number of factors that contribute to the performance of bilingual lexicon induction, including word frequency and burstiness, the size of the bilingual dictionary used for supervision, and the size of the comparable corpora. This chapter also includes a direct comparison with a previously state-of-the-art model of bilingual lexicon induction.
- Chapter 5 describes how we adapt our techniques for using comparable corpora to score the similarity between a pair of words in two languages to instead score a pair of multiword phrases. This chapter presents SMT experiments where we replace bilingually estimated feature functions with those estimated over comparable corpora.
- Chapter 6 applies our techniques for inducing translations (Chapter 4) and scoring a phrase table (Chapter 5) to end-to-end low resource SMT for several truly low resource source languages.

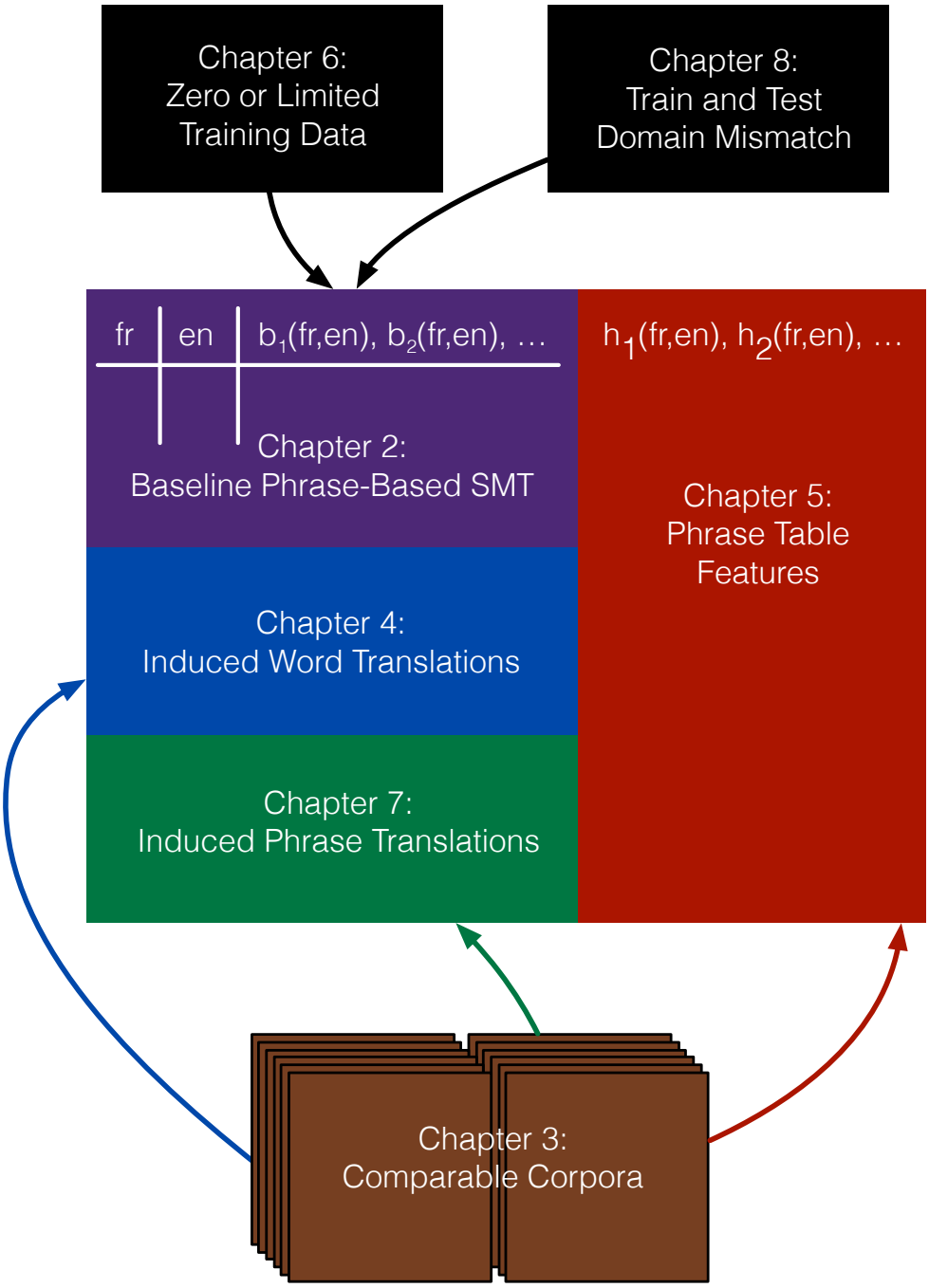


Figure 1.4: Schematic overview of the thesis. The large middle box depicts how we augment baseline phrase-based SMT models using comparable corpora. In Chapter 2, we define our baseline models, which are trained on parallel corpora. We describe our comparable corpora in Chapter 3. In Chapters 4 and 7, we present methods for using the comparable corpora to induce new source-target word and phrase translation pairs, respectively, which we use to augment baseline SMT models. In Chapter 5, we augment baseline models with new comparable corpora-based feature functions. In Chapters 6 and 8, we present end-to-end SMT experiments using comparable corpora to augment baseline models for translating low resource languages and domains.

- Chapter 7 presents our approach to inducing multiword phrase translations using a novel algorithm based on translation compositionality. It includes end-to-end SMT experiments that show that the quality of machine translations improve when we augment models with new phrasal translations in addition to unigram translations.
- Chapter 8 applies our bilingual lexicon induction and phrase table scoring techniques to a domain adaptation setting, where we have training data in some old-domain but wish to translate text in a new-domain. Additionally, this chapter presents an analysis of the errors that occur in the domain shift setting.
- Chapter 9 concludes the thesis with an overview of the major research findings.

1.3 Related publications

This thesis is based on seven publications:

- Chapters 3 and 8 present and extend "Measuring machine translation errors in new domains," which was published in the Transactions of the Association for Computational Linguistics (TACL) in 2013 and is joint work with John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu.
- Chapter 3 presents the data that we released along with "The Language Demographics of Amazon Mechanical Turk," which was published in the Transactions of the Association for Computational Linguistics (TACL) in 2014 and is joint work with Ellie Pavlick, Matt Post, Dmitry Kachaev, and Chris Callison-Burch.
- Chapter 4 extends "Supervised Bilingual Lexicon Induction with Multiple Monolingual Signals," which was published in the Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) in 2013.
- Chapter 5 presents "Toward Statistical Machine Translation without Parallel Corpora," which was published in the Proceedings of the Conference of the European Association for Computational Linguistics (EACL) in 2012 and is joint work with Alex Klementiev, Chris Callison-Burch, and David Yarowsky.
- Chapter 6 extends "Combining Bilingual and Comparable Corpora for Low Resource Machine Translation," which was published in the Proceedings of the Workshop on Statistical Machine Translation (WMT) in 2013.
- Chapter 7 elaborates on "Hallucinating Phrase Translations for Low Resource MT," which was published in the Proceedings of the Conference on Natural Language Learning (CoNLL) in 2014.
- Chapter 8 extends "Using Comparable Corpora to Adapt MT Models to New Domains," which was published in the Proceedings of the Workshop on Statistical Machine Translation (WMT) in 2014.

Chapter 2

Literature Review

2.1 Statistical Machine Translation

Statistical machine translation (SMT) was first formulated as a series of probabilistic models that learn word-to-word correspondences from sentence-aligned bilingual parallel corpora. IBM researchers (Brown *et al.*, 1988, 1993a) introduced and tested the first statistical machine translation (SMT) system trained on pairs of source and target sentence translations. They formulated SMT as a source-channel problem, as follows, where $\hat{e}$ is the optimal output translation, E is the set of all possible output sentences, and f is the input sentence:

$$\hat{e} = \underset{e \in E}{\operatorname{argmax}}\, p(e|f) = \underset{e \in E}{\operatorname{argmax}}\, [p(e) \cdot p(f|e)]$$

The translation model, $p(f|e)$ may be estimated by summing over all possible word alignments, A, in the corpus of translated sentence pairs:

$$\hat{e} = \underset{e \in E}{\operatorname{argmax}}\, p(e|f) = \underset{e \in E}{\operatorname{argmax}}\, [\sum_{a \in A} p(e) \cdot p(f, a|e)]$$

However, for efficiency, typically the single best word alignment is used to estimate the translation model parameters. Brown *et al.* (1990) proposed using the classic n-gram models often used in speech recognition to estimate the language model parameters, $p(e)$, which is the probability of a sequence of target language words:

$$p(e_1, e_2, e_3, ..., e_i) = p(e_1) \cdot p(e_2|e_1) \cdot p(e_3|e_1, e_2) \cdot ... \cdot p(e_i|e_1, e_2, e_3, ..., e_{i-1})$$

N-gram back off models estimate probabilities using local context, just $n - 1$ preceding words, rather than the full history. A trigram approximation is estimated as follows:

$$p(e_i|e_1, e_2, e_3, ..., e_{i-1}) \approx p(e_i|e_{i-1}, e_{i-2})$$

2.1.1 Phrase-Based Machine Translation

The major developments in the SMT field have built directly upon the initial IBM word-based statistical models. Marcu and Wong (2002) and Koehn *et al.* (2003) introduced phrase-based machine translation (PBMT), which takes advantage of multi-word phrase translations. Current methods, including both *phrase-based* (Koehn *et al.*, 2003; Och and Ney, 2002) and *hierarchical* models (Chiang, 2005), typically start by word-aligning a bilingual parallel corpus (Och and Ney, 2003). They extract multi-word phrases that are consistent with the Viterbi word alignments and use these phrases to build new translations. Phrase extraction heuristics (Och and Ney, 2004; Tillmann, 2003; Venugopal *et al.*, 2003) produce a set of phrase pairs (e, f), the phrase table, that are consistent with the word alignments.

Another important improvement to the original noisy channel formulation was proposed by Och and Ney (2002, 2004) and uses a linear model instead. This model was a crucial development in SMT because it allows the feature space to easily extend beyond basic translation and language model probabilities. It is formalized as:

$$\hat{e}_1^I = \underset{e_1^I \in E_1^I}{\operatorname{argmax}}\, \frac{1}{Z} \exp \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)$$

In this formulation, there are M features, each of which is a function, h_m, of the source and target strings. The source and target strings have lengths J and I, respectively. Features are typically defined for pairs of source and target phrases, and the score of a full target output sentence, e_1^I, is computed by summing over all of its sub-phrases. Z is a normalization constant, which can be ignored. During tuning, the feature weights, λ_m, are chosen to optimize performance on a development set of source and target sentence translation pairs. One widely used tuning method is Minimum Error Rate Training (MERT) (Och, 2003). Instead of maximum likelihood, MERT allows the optimization of objectives that measure translation quality. One such objective function is BLEU (Papineni *et al.*, 2002), which compares system translations to reference translations produced by people. The BLEU score is based on n-gram precision measures, which count how many n-grams in the reference translation are contained in the system output:

$$BLEU = BP \cdot exp(\sum_{n=1}^{N} w_n log p_n)$$

where BP is a brevity penalty, N is the maximum n-gram degree (typically 4), p_n is the precision over phrases of length n, and w_n is a weight associated with each precision metric (typically uniform, resulting in a geometric mean over n-gram precisions). We use BLEU to evaluate the quality machine translations throughout this thesis.

Recently, other tuning algorithms have been proposed that are more robust to a large feature space than MERT (Cherry and Foster, 2012; Chiang *et al.*, 2008, 2009; Hopkins and May, 2011). In this thesis, we make use of both MERT and the batch version of the Margin Infused Relaxed Algorithm (MIRA) (Crammer *et al.*, 2006). MIRA is an online learner that makes updates based on 'hope' hypothesis translations, which are both high quality (as measured by BLEU score) and reachable by the model, and 'fear' hypothesis translations, which are given high scores by the current model but are low quality translations (Chiang *et al.*, 2008). Cherry and Foster (2012) introduced a batch version of the online MIRA algorithm.

In PBMT, phrase table features typically include phrase translation probabilities, $\phi(e|f)$ and $\phi(f|e)$, which are calculated via maximum likelihood estimation over the word-aligned training corpus. Since MLE overestimates ϕ for phrase pairs with sparse counts, lexical weighting feature functions are used to smooth. Lexical weighting features consist of average word translation probabilities, $w(e_i|f_j)$, and are calculated via phrase-pair-internal word alignments (Koehn *et al.*, 2003). Other typical features are n-gram language model scores and a phrase penalty, which governs whether to use fewer longer phrases or more shorter phrases. Language model scores are estimated over target language monolingual corpora, and the phrase penalty feature is uniform across all phrase pairs.

Tillman (2004) define a "lexicalized reordering" model for PBMT, which is based upon the relative positions of target language phrases, which are translated according to the source language phrase sequence. That is, if two source phrases, s_{i-1} and s_i, translate into two target phrases, t_j and t_k, then a probability distribution is defined over the relative position of the target phrases. The phrase t_k can follow (be in order with) t_j, be just before it (swapped), or neither (discontinuous). Figure 2.1 shows these three reordering patterns. This reordering model is defined over the entire target sequence as follows:

$$P(t_1^n, o_1^n) \approx \prod_{i=1}^{n} p(t_i, o_i | t_{i-1}, o_{i-1}))$$

Where each o_i has three possible values: monotonic (in order), swapped, or discontinuous (neither). The probability of each phrase having each reordering value is dependent only upon the previously translated target phrase (t_{i-1}) and its reordering value.

The Moses SMT toolkit (Koehn *et al.*, 2007) contains an implementation of the now-standard PBMT pipeline, and we make use of it throughout this thesis. Figure 2.2 depicts the entire phrase-based SMT pipeline.

2.1.2 Other Models of Translation

Syntax-Based SMT Syntax-based machine translation grammars have shown improvements over phrase-based grammars on some language pairs (Galley *et al.*, 2004; Huang and Mi, 2010; Li *et al.*, 2010b; Zollmann and Venugopal, 2006). There are essentially two ways to incorporate a syntax model into an SMT system. The first is to extract grammar rules which have some syntactic structure and which combine into complete parse trees on the source side, target side, or both. This is approach is enabled by not only high-performance parsers but also improvements to syntactic decoding (Galley *et al.*, 2006; Huang and Mi, 2010; Li *et al.*, 2010a; Quirk and Menezes, 2006; Quirk *et al.*, 2005;

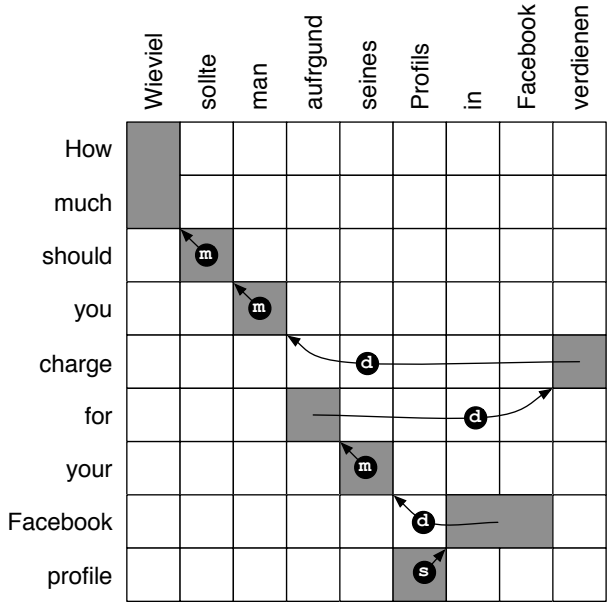


Figure 2.1: The reordering probabilities from the phrase-based models are estimated from bilingual data by calculating how often in the parallel corpus a phrase pair (f, e) is orientated with the preceding phrase pair in the 3 types of orientations (**m**onotone, **s**wapped, and **d**iscontinuous).

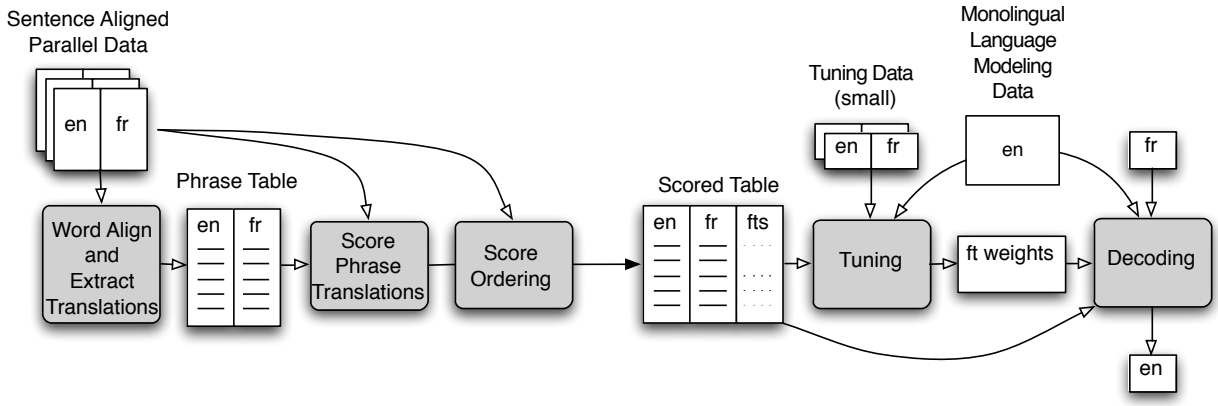


Figure 2.2: Standard statistical machine translation pipeline. Parallel data is automatically word aligned. Using the word alignments, a phrase table is extracted and scored. Discriminative training uses a small amount of parallel data to set the weights for each parameter in the scored phrase table as well as for the language model, which is trained on monolingual target language text. Finally, the scored tables, language model, and learned weights are used to decode new text.

Zollmann and Venugopal, 2006). Somewhat complementary, Koehn and Hoang (2007) integrates syntactic and morphological information into grammar rules at the *word* level. The second way to integrate this information is to add syntactically-informed features to any set of grammar rules, which may or may not contain syntactic structure (e.g., Hanneman *et al.* (2010)). This approach is enabled by recent developments in methods to efficiently and accurately tune a large number of SMT feature parameters (Cherry and Foster, 2012; Chiang *et al.*, 2009; Hopkins and May, 2011; Liang *et al.*, 2006).

Decipherment Decipherment combines techniques from cryptography and machine translation. It is typically cast as an unsupervised substitution and transposition problem or, comparably, the task of building a translation model when there is zero parallel text available to learn from. Knight *et al.* (2006) explore character and phonetic ciphers as well as a toy word-substitution cipher. That work uses expectation-maximization to learn a noisy-channel model and implements several slight modifications to improve performance. Snyder *et al.* (2010) discover word translations for a lost language by mapping them onto cognates of a known related, modern language. That work only discovers a bilingual lexicon and does not attempt to translate word sequences.

Ravi and Knight (2011) go further and use decipherment to translate multi-word expressions, and Nuhn *et al.* (2012) and Ravi (2013) use pruning based on context similarity and a hash sampling strategy, respectively, to improve training efficiency. Dou and Knight (2012) use decipherment over adjacent bigrams to learn new word translations. That work focuses on a domain adaptation setting and augments an old-domain SMT model with new word translations deciphered from comparable (nearly parallel) new-domain texts. Dou and Knight (2013) move from adjacent bigrams to dependency bigrams. In that work, again, the new, deciphered lexicon is used to augment and adapt an old-domain SMT model to a new-domain of text. In contrast to decipherment, we use small amounts of parallel data to train supervised models, which we then use to augment baseline SMT systems.

2.1.3 Low Resource Machine Translation

Because the standard pipeline for training SMT models relies exclusively on parallel text, when such data is not available in large quantities, machine translation quality tends to be poor. Much of the prior work on low resource machine translation has focused on individual pipeline components or error types. For example, Xiang *et al.* (2010) presents an alignment combination-based method for doing unsupervised word alignment on small amounts of bitext, improving the word alignment component of the SMT pipeline. As we show in Chapter 3, errors due to unknown words are a major reason for the performance degradation in low resource conditions, and a lot of prior research has focused on these errors. We review that work in Section 2.1.4.

In addition to such focused efforts, there have been several long-term projects on low resource SMT that have focused on improving the entire pipeline for particular language pairs. Much of this work relies on language-specific resources and linguistic analysis. Here, we describe two of the major, recent efforts in developing high quality MT systems for particular low resource language pairs. In contrast to these projects, we take a language-independent approach to low resource machine translation in this thesis.

2.1.3.1 AVENUE

In the early 2000s, the AVENUE project (Carbonell *et al.*, 2002; Lavie *et al.*, 2003; Probst *et al.*, 2002) researched ways to rapidly develop MT systems for low-resource languages. The project's first major MT system learned so-called transfer-rules, which are synchronous context free grammar (SCFG) rules that also have the following: explicit alignments between components (terminals or non-terminals), constraints on the source side, constraints on the target side, and source-target constraints which dictate what features are transferred from the source to the target when the rule is applied. Rules can be manually or automatically specified. In Probst *et al.* (2001), the rules are learned from a carefully controlled corpus, which is intended to include examples of all basic linguistic structures and minimal pairs which allow for linguistic feature detection [1] (Probst *et al.*, 2001), and in Lavie *et al.* (2003) they are learned from a combination of controlled and naturally occurring corpora. The assumption is that learning at least some transfer rules from a controlled, linguistically representative corpus will yield better MT performance than a random set of translated sentences. Automatically learned transfer rules are lexically tied to the training corpus but are later generalized and composed. Probst (2003) describes how a bilingual corpus can be used to project target (high resource) side POS

[1] For example, the minimal pairs *the rock fell* and *the rocks fell* are both presented in order to determine if the low resource language marks singular and plural nouns and the corresponding verbs differently. If a language does distinguish plural from singular, the elicitation process will go on to determine if it also distinguishes dual nouns. It takes advantage of well-known typological implications to limit the search. For example, if a language does not distinguish plural from singular nouns, it will not distinguish dual.

and morphological tags onto a source (low resource) side lexicon. Lavie *et al.* (2004) extends the AVENUE model to include frequency-based probabilities on the ruleset and does decoding over a translation lattice.

More recently, Monson *et al.* (2008) explains specific extensions of the original AVENUE system. In addition to the integration of language-dependent paradigm-based morphological analyzers, that work describes a top down SCFG grammar extractor. Llitjós (2006) describes how to incorporate user feedback into the model, and Alvarez *et al.* (2006) describes the Minor Language Elicitation (MILE) corpus, which is a set of about 12,000 English sentences selected in order to demonstrate a comprehensive set of linguistic phenomena and which the paper claims is a good corpus to have translated into low resource languages. Clark *et al.* (2008b) use the Urdu translation of the MILE corpus to evaluate their system for detecting linguistic typological features, which is based upon the dataset's detailed annotations and a method for clustering sentence minimal pairs. Clark *et al.* (2008a) extends this idea and develops an active learning framework for choosing sentences from the MILE corpus to have annotated next, based upon the features detected so far. That work uses the World Atlas of Language Structures (Haspelmath *et al.*, 2005) for evaluation and the feature detection system is reported to detect the correct value for 18 of the 21 target typological features for Spanish.

2.1.3.2 METIS-II

Carl *et al.* (2008), Vandeghinste *et al.* (2008), and Carl (2009) present the METIS-II low resource[2] MT system. This statistical MT system requires neither hand-written translation rules nor any parallel text. However, it does require POS tagging and some morphological analysis on the low-resource source side text and performance increases if parses are also available. Essentially, translation is done through the use of a POS-tagged lemmatized dictionary. The decoder uses morphological information to choose among English full word forms, and, in some cases, and reordering is done through language-dependent rules. Although this system does not depend on parallel text, it does heavily depend upon sophisticated language-dependent resources.

2.1.3.3 Other work

Gangadharaiah (2011) tackles several data sparsity issues within the example-based machine translation (EBMT) framework. These include translating OOV and rare words not only to increase the coverage of the translation model, but also that of the target side language model. She also develops ways to automatically cluster translation phrase pairs. The thesis also explores using phrase and phrase pair clusters to define the slots in translation and language modeling templates, or class-based models. This work attempts to tackle some of the same data sparsity issues that our work proposes to handle including, in particular, phrase table coverage. However, our models for doing so are quite different and focus much more on the use of a variety of new non-parallel data resources. Laukaitis and Vasilecas (2007) describes a hybrid approach to building an MT system to translate from English into a low resource language (Lithuanian). Translating from English, this work is able to do a very deep linguistic analysis of source side text.

2.1.4 The OOV and rare word problem in SMT

Out-of-vocabulary (OOV) words, or source language words never seen during training but that then appear at test time, is a well-acknowledged challenge for SMT, especially in low resource conditions. When an SMT system has no evidence for how to translate these new words, they are typically either deleted or copied into the output identically, which negatively impacts the quality of output translations in a very obvious way. The most common methods for dealing with such lexical items include morphological analysis (Popovic and Ney, 2004; Yang and Kirchhoff, 2006), stemming, using synonym lists (Callison-Burch *et al.*, 2006), transliteration (Hassan and Sorensen, 2005; Irvine *et al.*, 2010a; Vilar *et al.*, 2007), spelling expansion/correction, dictionary expansion/integration (Okuma *et al.*, 2007), or some combination of these (Habash, 2008).

Carbonell *et al.* (2006) and Gangadharaiah *et al.* (2010) present an approach to translate rare and OOV source language words identically onto the target language and then use target language distributional similarity metrics and a very large target side language model to identify translations. Callison-Burch *et al.* (2006) and Marton *et al.* (2009) use monolingually-derived paraphrases to directly expand the SMT phrase table coverage. That is, that work identifies OOV phrases in source language text and uses a large source language monolingual corpus to identify one or more paraphrases for each OOV phrase. Existing phrase table entries for the paraphrases can then be used to add entries for the original OOV source phrase. Somewhat similarly, Talbot and Osborne (2006) cluster 'lexically redundant'

[2]This group refers to languages for which there are few NLP resource and datasets as 'small languages.'

source word types in order to reduce data sparsity. Other prior research (e.g., Goldwater and McClosky (2005); Luong *et al.* (2010); Virpioja *et al.* (2010)) has moved the SMT unit of analysis from the *word* to the *morpheme* or even character substrings (Neubig *et al.*, 2012). Such approaches have the potential to reduce the number of unknown source words due to rich morphologies. Nakov and Ng (2011) tackle morphologically rich source languages by treating morphologically similar words as potential paraphrases.

While some approaches to the OOV problem (e.g., morphological analysis, stemming, paraphrasing) attempt to map the OOV word onto an in-vocabulary source word, others (e.g., transliteration, dictionary expansion) output target language word(s) directly. In this thesis, we focus on solving the OOV problem by identifying target language translations directly. In Section 2.2.1, we review relevant literature on bilingual lexicon induction, which is one dictionary expansion method that is often approached as a separate task.

2.2 Expanding bilingual resources

As we showed in Chapter 1, the success of statistical machine translation is dependent on the amount and type of parallel training data that is used to estimate models. Thus, prior research has focused on expanding the bilingual resources available for training. Here, we review four approaches. The first two use *comparable corpora* to identify additional pairs of translated words or sentences. Comparable corpora are pairs of monolingual corpora that have some overlap, for example in genre or topic. In Section 3.3, we describe the notion of comparable corpora in more detail as well as those that we use and release in this thesis. The third approach to expanding bilingual resources crowdsources manual annotations, which has been made possible largely due to the Amazon Mechanical Turk platform (Callison-Burch and Dredze, 2010). Finally, we review a variety of methods for automatically expanding the coverage of SMT models directly.

2.2.1 Bilingual lexical induction

Bilingual lexicon induction is the task of identifying pairs of word translations from monolingual or comparable corpora. Additionally, a small seed dictionary is also typically assumed. Induced word translations can be used to expand the coverage of translation models extracted from parallel corpora, for example to translate OOV words (Section 2.1.4). However, most prior work has treated bilingual lexicon induction as a standalone task, without actually integrating induced translations into end-to-end machine translation. In this thesis, we present a novel approach to doing bilingual lexicon induction in Chapter 4 and use then induced translations in end-to-end SMT in Chapters 6 and 8. Here, we review prior work on bilingual lexicon induction. We give an overview of early work on contextual similarity and then briefly cover a plethora of other approaches to the task and, finally, describe work on integrating induced translations into end-to-end SMT.

2.2.1.1 Contextual Similarity

Rapp (1995) and Fung (1995) were the first to propose using the context of a given word as a clue to its translation. Rapp (1995) creates co-occurrence matrices for the source and target languages, where the co-occurrence between a pair of words is defined as follows:

$$A_{i,j} = \frac{(f(i,j))^2}{f(i) \cdot f(j)}$$

Where $f(i,j)$ is defined as the number of times words i and j, in the same language, occur in the same context in a large monolingual corpus (Rapp (1995) uses a context window of 11 words), and $f(i)$ is the total number of times word i appears in the same corpus. After computing the two co-occurrence matrices, that work iteratively randomly permutes one of them and calculates the similarity between them. The permutation is optimal when the similarity between the matrices is maximal, which is when the ordered words in the two matrices are most likely to be translations of one another. Results are given for a set of 100 English and German word translation pairs.

Later work, including Fung and Yee (1998) and Rapp (1999), use small seed dictionaries to *project word-based context vectors* from one language into the other. That is, each position in contextual vector v corresponds to a word in the source vocabulary[3], and vectors v are computed for each source word in the test set. Fung and Yee (1998) calculates

[3] In fact, they need only correspond to those source words which have translations in the seed bilingual dictionary.

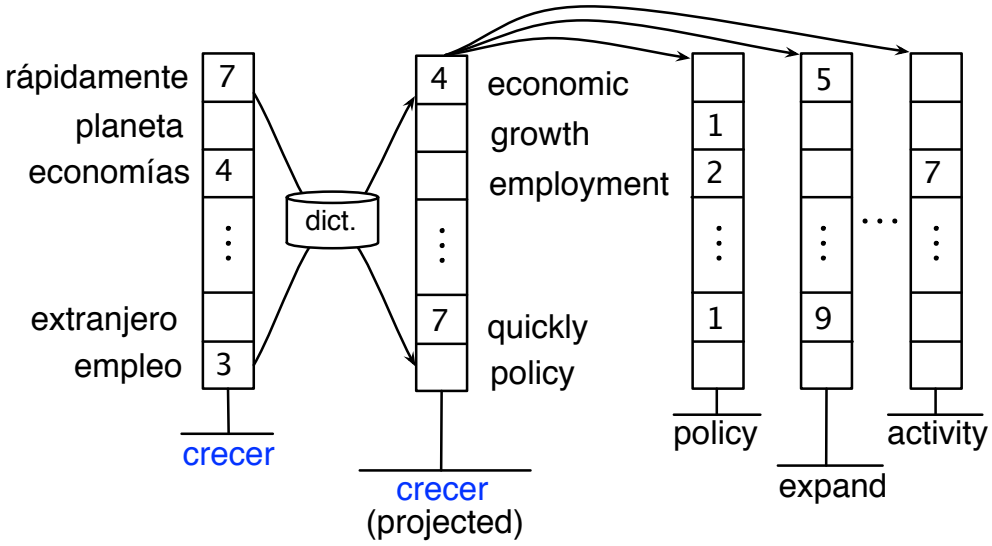


Figure 2.3: Example of projecting contextual vectors over a seed bilingual lexicon. In monolingual text, Spanish *crecer* appears in the context of the words *empleo, extranjero*, etc. A context vector is built and projected across a seed dictionary. Context vectors for English words (*policy, expand*, etc.) are collected and then compared against the projected context vector for Spanish *crecer*. Words with similar context vectors are likely to be translations of one another.

the *i*th position of word w's context vector, v_{w_i}, as:

$$v_{w_i} = TF_{i,w} \cdot IDF_i$$

Where $TF_{i,w}$ is the number of times i and w co-occur (in this case, defined as appearing in the same sentence), and:

$$IDF_i = log \frac{maxn}{f_i} + 1$$

Where $maxn$ is the maximum frequency of any word in the corpus, and f_i is the frequency of word i. Rapp (1999) uses log-likelihood ratios instead of $TF \cdot IDF$. Once source and target language contextual vectors are built, each position in the source language vectors is projected onto the target side using a seed bilingual dictionary. Finally, *contextual similarities* are calculated. That is, each projected vector is compared, using any vector comparison method, with the context vector of each target word. Word pairs with high contextual similarity are likely to be translations. This method of projecting contextual vectors is illustrated in Figure 2.3. Rapp (1999) uses the same projection method as Fung and Yee (1998) but uses log-likelihood ratios instead of $TF \cdot IDF$.

Other work has used dependency relations in place of adjacent words to define context (Andrade *et al.*, 2012; Garera *et al.*, 2009). Turney and Pantel (2010) give a thorough overview of vector space models of meaning.

2.2.1.2 Other Monolingual Similarity Metrics

Schafer and Yarowsky (2002) exploit the idea that word translations tend to co-occur in time across languages, and Schafer (2006) uses this and other contextual similarity measures to bootstrap a small seed bilingual dictionary and induce full dictionaries for low resource languages. Klementiev and Roth (2006) also use the temporal cue to train a phonetic similarity model for associating Named Entities across languages. Koehn and Knight (2002) use similarity in spelling as another kind of cue that a pair of words may be translations of one another. Haghighi *et al.* (2008) make use of contextual and orthographic clues to learn a generative model from monolingual texts and a seed lexicon. We provide further details and compare directly against the model proposed by Haghighi *et al.* (2008) in Section 4.6.

Recent work has used graph-based models to induce translations. Mausam *et al.* (2010) uses freely available online dictionaries and inference over translation graphs to compile a very large, multilingual dictionary. Laws *et al.* (2010) use graph-based models to represent linguistic relations and induce translations. Tamura *et al.* (2012) employ the classic notions of co-occurrence and contextual similarity but use graph-based label propagation to induce translations.

Approaching the problem from an information retrieval perspective, Zhang *et al.* (2005) use a system based on cross-lingual query expansion to identify translations for OOV words. Other recent work includes Mimno *et al.* (2009), which proposes a polylingual topic model and matched high probability words in each topic across languages.

2.2.1.3 Integration with SMT

All bilingual lexicon induction and dictionary expansion methods could be used to supplement parallel data used for estimating word alignments and scored phrase tables. The most obvious way to integrate lexicon induction output into the SMT pipeline would be to induce translations for out-of-vocabulary and rare words. That is, if a word in our test set does not have a translation in the phrase table, we could induce one for it. Although most work on bilingual lexicon induction is motivated by the idea that outputs could be integrated into end-to-end SMT, until recently such an extrinsic evaluation was rarely performed. Daumé and Jagarlamudi (2011) use canonical correlation analysis (CCA) and both contextual and orthographic features to induce translations. Razmara *et al.* (2013) construct a graph using source language monolingual text and identify translations for source language OOV words by pivoting through paraphrases. In Irvine *et al.* (2013b), we presented a method for expanding an initial translation dictionary estimated from old-domain parallel corpora by matching marginal probabilities over new-domain comparable corpora. Daumé and Jagarlamudi (2011), Razmara *et al.* (2013), and our prior work in Irvine *et al.* (2013b) integrate translations into an SMT model to improve performance in domain adaptation settings.

2.2.2 Extracting Parallel Data from Comparable Corpora

Resnik and Smith (2003), Munteanu and Marcu (2005), Abdul-Rauf and Schwenk (2009a), Abdul-Rauf and Schwenk (2009b), and Smith *et al.* (2010) identify parallel sentences in comparable corpora, and Munteanu and Marcu (2006) extend that original work to identify parallel sub-sentential fragments. The latter uses a probabilistic lexicon and information retrieval methods to identify similar document pairs and then uses the same word translation probabilities to detect parallel fragments within the document pairs. They supplement exiting parallel data with the new sentence and fragment pairs evaluate end-to-end SMT systems trained on the augmented parallel datasets. Quirk *et al.* (2007) also seek to identify phrase translation pairs from comparable corpora, but that method requires a first pass identification of "promising" comparable pairs of sentences from paired comparable documents. It then uses a generative model to extract fragment translation pairs. Similarly, Hewavitharana and Vogel (2011) seek to identify phrase translation pairs from comparable corpora but require a first pass to identify a set of comparable sentences and then a second pass through the data to find the best phrasal alignment within each sentence pair. These efforts at using comparable corpora to expand parallel corpora are orthogonal to the approaches that we propose in this thesis. We use parallel corpora, when available, and use comparable corpora to augment SMT models without assuming that they contain any novel parallel text.

2.2.3 Crowdsourcing Translations

New online services like Amazon's Mechanical Turk[4] (MTurk) have made it possible to elicit translations from native speakers at a relatively low cost. MTurk is an online platform where 'requesters' can pay 'workers' small amounts of money to complete Human Intelligence Tasks (HITs). MTurk has been used in a variety of NLP research (e.g. Callison-Burch and Dredze (2010); Snow *et al.* (2008); Zaidan and Callison-Burch (2011)). In our own work, we have used MTurk to compile large bilingual dictionaries (Irvine and Klementiev, 2010; Pavlick *et al.*, 2014). In Section 3.2.2, we describe our crowdsourced dictionaries in detail. Zaidan and Callison-Burch (2011) spent less than $1,500 to recreate the NIST Urdu Evaluation set, which consists of four English references for each of nearly 2,000 Urdu sentences. Post *et al.* (2012) used MTurk to construct training, tuning, and testing parallel datasets for six low resource Indian languages. Carbonell (2010) discusses some of the issues in collecting resources for low resource languages with the goal of developing MT systems. He suggests that collecting treebanks, comprehensive bilingual dictionaries, and detailed linguistic analyses may provide more gains than simply supplementing the available parallel data. However, in the case of non-expert annotators, we cannot expect to be able to gather such thorough linguistic analyses. Ambati and Carbonell (2009) introduce proactive learning, which could help select sentences and annotation types that would most cost-effectively increase the performance of an MT system for a low-resource language.

[4] www.mturk.com

2.2.4 Automatically Expanding SMT Model Coverage

Instead of expanding general bilingual resources from which we can extract or build SMT models, some work has augmented baseline models with synthetic phrase translations which are proposed automatically. In Chapter 7, we present such a method based on phrase translation composition in order to expand the coverage of models trained in low resource conditions.

Chahuneau *et al.* (2013) generate new synthetic phrase translations containing new target side morphological inflections. Ammar *et al.* (2013) propose new phrase pairs containing transliterations, inserted function words, and morphological inflections. Tsvetkov *et al.* (2013) vary determiners in order to generate synthetic phrase pairs, and Tsvetkov *et al.* (2014) use commonly confused phones for spoken language translation.

Sharoff *et al.* (2006) and Babych *et al.* (2007) use a combination of filtering and composing multiword translations that is similar the the approach we propose in Chapter 7. However, they use their model to propose translations of difficult expressions to human translators. Also somewhat similar to our approach, Garera and Yarowsky (2008) compose translations of compound nouns using known unigram noun translations and pivoting across many languages.[5] However, unlike that work, our method induces translations for any arbitrary multi-word phrase, not just noun-noun compounds, and we use the translations in end-to-end SMT.

2.3 Domain Adaptation for Machine Translation

Domain adaptation is the task of modifying models trained on text in some old domain to work well on text from a new domain. Prior work has adapted part-of-speech taggers (Blitzer *et al.*, 2006; Daumé, 2007) and sentiment classifiers (Blitzer *et al.*, 2007), for example, to new domains of text. Building on the successes of domain adaptation in general natural language processing and machine learning settings, domain adaptation for machine translation is a rapidly growing area of research. The 'domain' of a text is typically assumed to be given, as Sekine (1997) states, 'artificially or intuitively defined by a human' (p. 2). Examples of contrasting 'domains' range from topical (e.g. reviews of books versus reviews of kitchen appliances (Blitzer *et al.*, 2007)) to channel (e.g. newswire versus twitter (Gimpel *et al.*, 2011)) to genre or (e.g. speeches versus press releases (Foster and Kuhn, 2007)). Kilgarriff and Rose (1998) present several information theoretic measures to estimate the similarity between a pair of documents, which could be used to measure the homogeneity of a hypothesis domain. However, the question of how we should define and identify text domains most effectively for natural language processing applications remains open. In this thesis, like the vast majority of prior NLP research, we assume that domain is predefined.

Recent work on machine translation domain adaptation has focused on either the language modeling component or the translation modeling component of an SMT model. Language modeling research has explored methods for subselecting new-domain data from a large monolingual target language corpus for use as language model training data (Gao *et al.*, 2002; Klakow, 2000; Lin *et al.*, 1997; Mansour *et al.*, 2011; Moore and Lewis, 2010). Translation modeling research has typically assumed that either (1) two parallel datasets are available, one in the old domain and one in the new, or (2) a large, mixed-domain parallel training corpus is available. In the first setting, the goal is to effectively make use of both the old-domain and the new-domain parallel training corpora (Civera and Juan, 2007; Foster and Kuhn, 2007; Foster *et al.*, 2010; Haddow, 2013; Haddow and Koehn, 2012; Koehn and Schroeder, 2007). In the second setting, it has been shown that, in some cases, training a translation model on a subset of new-domain parallel training data within a larger training corpus can be more effective than using the complete dataset (Axelrod *et al.*, 2011; Gascó *et al.*, 2012; Mansour *et al.*, 2011; Sennrich, 2012).

For many language pairs and domains, *no* new-domain parallel training data is available. Wu *et al.* (2008) machine translate new-domain source language monolingual corpora and use the synthetic parallel corpus as additional training data. Daumé and Jagarlamudi (2011), Zhang and Zong (2013), as well as our prior work in Irvine *et al.* (2013b) use new-domain comparable corpora to mine translations for unseen words. That work follows a long line of research on bilingual lexicon induction (Section 2.2.1). These efforts improve errors due to unseen translations. Our work in Irvine and Callison-Burch (2014b) (Chapter 8) is the first to focus on fixing errors due to inaccurate translation model *scores* in the setting where no new-domain parallel training data is available.

[5] e.g. Albanian *hekurudhë* → *hekur udhë* splitting using lexicon lookup, → English *iron path* using unigram translations, → Italian *ferro via*, German *eisen bahn*, and Swedish *yärn väg* using unigram translations into one or many *other* languages, → *ferrovia*, *eisenbahn*, *yärnväg* by simple concatenation, → English *railway* using known translations

Chapter 3

Languages, Data, and Analysis

As we showed in Chapter 1, the amount and type of textual data resources available for a given language pair has substantial impact on the performance of the SMT model that we are able to estimate. This chapter provides an overview of publicly available datasets for training SMT models for a large number of languages paired with English. First, we enumerate all of the languages that we study in the thesis. Then, in Section 3.2, we describe parallel datasets and dictionaries. We introduce monolingual and comparable corpora and those corpora that we have harvested from the web in Section 3.3. Throughout the thesis, we use these corpora to augment traditionally trained models of translation. Finally, in 3.4, we present results from a novel error analysis technique to better understand what goes wrong in machine translation when we train models in low resource conditions. We include an analysis of word alignment errors and introduce a set of manual word alignments.

One major contribution of this thesis is the release of language packs for 151 languages paired with English. Each language pack includes the following:

1. Bilingual dictionary of source language words and their English translations (96 of the 151 languages; described in Section 3.2.2)

2. Comparable corpus of time-stamped source language online news text paired with English news text collected on the same dates (115 of the 151 languages; described in Section 3.3.1).

3. Comparable corpus of source language Wikipedia pages paired with their inter-lingually linked English counterparts (142 of the 151 languages; described in Section 3.3.2).

4. For non-roman script languages: a dictionary of source language names paired with English equivalents, which are likely transliterations (described in Section 3.3.2.).

3.1 Languages

Our set of languages of interest include both truly low resource languages (e.g. Somali, Nepali, Kyrgyz, Tigrinya) and high resource languages (e.g. French, Spanish, Arabic, Chinese). Additionally, our language set includes a large degree of linguistic diversity. Table 10.1 in Appendix 10 shows some basic linguistic features for each language as reported in the World Atlas of Language Structures (WALS) (Haspelmath *et al.*, 2005). Of the 151 languages, 78 are from the Indo-European language family, 17 are Austronesian, 11 Altaic, 7 Afro-Asiatic, 7 Niger-Congo, 5 Sino-Tibetan, 4 each are from the Dravidian and Uralic families, and 18 are from a variety of other families, including Quechuan, Eskimo-Alteut, Tai-Kadai, and constructed languages, among others. Twenty-one of the languages are spoken in India; seven are spoken in Spain, seven in the Philippines, six in Pakistan, six in Switzerland, five in France, and four in Indonesia. Table 10.1 also presents the canonical sentential word order for each language; 47 are subject-verb-object (SVO) ordered and 43 are subject-object-verb (SOV). Additionally, eight are VSO, two are VOS (Gilbertese and Malagasy, both Austronesian languages), and eight have no dominant word order. The word order of the remaining languages is not reported in WALS. Sixty-three languages are strongly suffixing, while another eleven are weakly suffixing. Fifteen languages have little affixation while only three are strongly prefixing (Luganda, Shona, and Zulu, which are all Niger-Congo languages). Most prefixing languages are either spoken in the southern half of Africa (e.g. Bantu languages)

or are native North American and Mesoamerican languages, and currently little electronic textual data is available for these language groups (Dryer, 2011).

3.2 Parallel Corpora and Dictionaries

3.2.1 Parallel Corpora

As we described in Chapter 2, Gale and Church (1991, 1993) and Brown *et al.* (1993b) first proposed the idea of using pairs of translated sentences, or a parallel corpus, to train a statistical machine translation model. Since then, standard approaches have moved from word-based models to phrase-based models (Koehn *et al.*, 2003) and, more recently, to syntax-based models (Chiang, 2005; Galley *et al.*, 2004; Zollmann and Venugopal, 2006). All these standard approaches to statistical machine translation still estimate model parameters exclusively from parallel corpora.

The largest publicly available parallel corpus is the French-English 10^9 corpus, which contains nearly 1 billion words of sentence aligned French and English extracted automatically from a variety of websites (Callison-Burch *et al.*, 2009). Additionally, the French-English Hansard Canadian parliamentary proceedings[1] contain about 150 million word tokens of each language. As part of the DARPA Global Autonomous Language Exploitation (GALE) program, about 200 million words of Chinese-English and Arabic-English parallel text was compiled and made available for research. The Europarl parallel corpus (Koehn, 2005) is another large, commonly used SMT dataset. It contains text in 21 languages extracted from the proceedings of the European Parliament. The corpora range in size from about 50 million words in Danish, German, Spanish, Finnish, French, Italian, Dutch, Portuguese, and Swedish paired with English to about 10 million words for Bulgarian and Romanian. Similarly, the Acquis corpus contains 4-8 million words of parallel EU legal documents in each of 21 EU languages (Steinberger *et al.*, 2006).

The MultiUN corpus[2] is a collection of parallel text published on the United Nations website between 2000 and 2009 (Eisele and Chen, 2010). The Common Crawl parallel corpus (Smith *et al.*, 2013) is a collection of parallel texts in 18 languages paired with English. That dataset was harvested automatically from the Common Crawl corpus, a 102TB snapshot of about 2 billion public webpages.[3] The corpora range from 130 million words of French-English text to 200 thousand words of Pashto-English. Other low resource languages included in the Common Crawl parallel corpus are Kannada, Somali, Telugu, Farsi, Bengali, Urdu, and Tamil. One million words of text or fewer is available for each. Several datasets have been a byproduct of the annual Workshop on Statistical Machine Translation,[4] including the News Commentary corpus, which consists of about 3 million words of news text and commentary translated from English into Czech, German, Spanish, French, and Russian. In this thesis, we also make use of parallel corpora compiled through the use of crowdsourcing on Amazon's Mechanical Turk platform (Post *et al.*, 2012).

A variety of other parallel corpora have been released, many through the Open Parallel Corpus (OPUS) project[5] and the European Language Resources Association (ELRA).[6] Many of the OPUS and ELRA bitexts contain text from a particular domain, such as movie subtitles and biomedical data (mostly prescription drug labels) from the European Medicines Agency (EMEA). In Chapter 8, we address methods for adapting SMT models to new text domains.

A summary of the availability of parallel text for 96 languages paired with English is given in Figure 3.1 and the corresponding values are given in Table 11.1 in Appendix 11. Only languages for which some parallel data is available are plotted.[7] Estimates of the number of speakers of each language are taken from the Ethnologue[8] and are shown on the x-axis. The amount of parallel data is given on the y-axis. It is important to keep in mind that the diversity of content and translation quality vary dramatically across corpora. For example, the OPUS EMEA corpora contain small vocabularies and large numbers of duplicate phrases, giving them low value for training models to translate text outside of the narrow drug prescription labels domain.

Figure 3.1 shows a positive correlation (Pearson's $r = 0.58$ on raw, non-log transformed data) between the number of speakers and amount of available parallel text. For example, Chinese and Spanish have the largest number of speakers and very large parallel corpora are available for both. Similarly, Maori and Gaelic have few speakers and very

[1] www.parl.gc.ca
[2] www.euromatrixplus.net/multi-un/
[3] commoncrawl.org/
[4] www.statmt.org/
[5] opus.lingfil.uu.se/
[6] www.elra.info/
[7] We have not included the Bible or Book of Mormon in our estimates, which are both available for nearly all languages in the world.
[8] www.ethnologue.com/

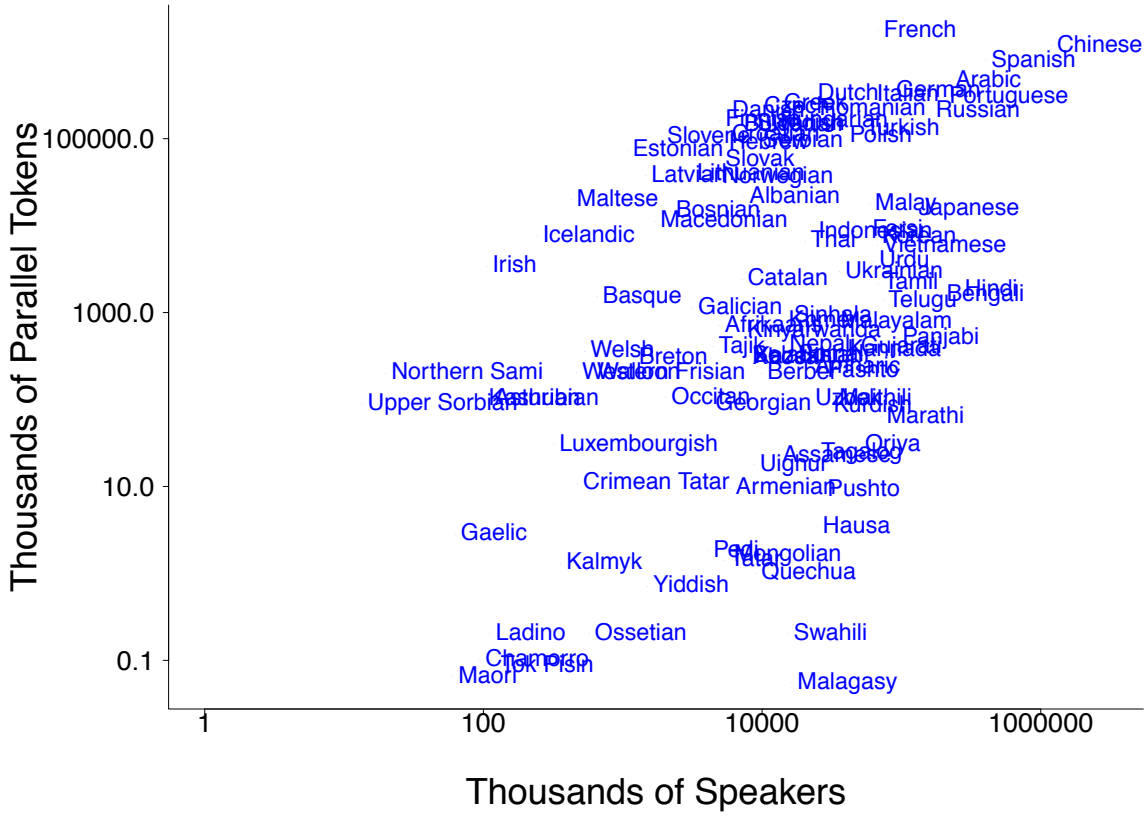


Figure 3.1: Thousands of speakers by millions of publicly available words of parallel data

little parallel data. In contrast, there are 260 and 190 million Hindi and Bengali speakers, respectively, but less than 2 million words of parallel data is available for each.

3.2.2 Bilingual Dictionaries

Throughout the thesis we also make use of bilingual dictionaries, which consist of translations of words and, in a few instances, short phrases. We use (1) dictionaries extracted from word aligned parallel data, (2) existing electronic dictionaries, (3) paper dictionaries scanned and digitized with optical character recognition (OCR),[9] and (4) crowdsourced dictionaries, gathered via Amazon Mechanical Turk (MTurk). Table 11.2 in Appendix 11 gives data statistics for all of the dictionaries except those extracted from word-aligned parallel corpora, which vary across experiments.

In Pavlick *et al.* (2014), we describe an empirical study of the languages spoken by workers on Mechanical Turk. In that work, we focused on the 100 languages which have the largest number of Wikipedia articles and posted HITs asking workers to translate the most frequent $10,000$ words in the most viewed $1,000$ pages for each source language. Although all of the source words in the Wikipedia dictionaries are unigrams, we allowed workers to translate them into multi-word English phrases. Workers were shown words in the context of three Wikipedia sentences. Additional details on experimental design and quality control mechanisms are given in Pavlick *et al.* (2014). As a result of that project, for languages that do have high coverage on MTurk, we collected bilingual dictionaries of about $10,000$ words translated into English. For the purposes of experiments in this thesis, we filter the dictionaries to include only high quality translations. Specifically, we only use translations that have a quality score of at least 0.6 under the metric given by Pavlick *et al.* (2014).

[9] Thanks to David Yarowsky for sharing his large collections of scanned dictionaries.

3.3 Monolingual and Comparable Corpora

In his thesis, Schafer (2006) presents a detailed hierarchy of text resources used in machine translation, sorted by the amount of supervision included in each. He describes the most supervised resources to be aligned source and target language treebanks and the least supervised resources to be independent, monolingual source and target language corpora without accompanying bilingual dictionaries. Monolingual corpora may be comparable in genre, topic, and/or associated timestamps. In this thesis, we refer to monolingual corpora which overlap in genre as well as either topic or associate timestamps as *comparable* corpora. Prior work has attempted to measure the degree of comparability between a pair of corpora, for example using an existing bilingual dictionary (Li and Gaussier, 2010; Li *et al.*, 2011) or using performance on some task as a proxy (Su and Babych, 2012). We do not attempt to measure the comparability between our corpora but, rather, we select corpora such that they are likely to contain text on the same topics. We use two sources of monolingual comparable text corpora: (1) web crawls of online newspapers, and (2) Wikipedia. In both cases, we harvest foreign language text as well as English text and estimate document alignments to link our comparable corpora.

3.3.1 Web crawls

Online newspapers are good sources of high quality text for many languages. We began harvesting such data by crawling several well-known news sources that publish stories in two or more languages, including Deutsch Welle and Voice of America. In order to gather more data, particularly for less commonly used languages, we scraped a list of $44,892$ newspapers and their locations, URLs, and languages from the ABYZ News Links website.[10] The resulting database of newspapers contains links to online newspapers published in 128 languages, and we set up web crawls to download the content from each daily.

Although ABYZ News Links provides labels of the primary publication language for each newspaper, we found that in some cases these labels were incorrect and, in other cases, newspapers contain content in multiple languages. This is particularly true for publications in regions where several languages are spoken. In order to obtain cleaner labels, we use two language identification models: (1) the Compact Language Detection 2 (CLD2) system,[11] and (2) the system released by Bergsma *et al.* (2012). The CLD2 model consists of pre-trained Naive Bayes classifiers for over 80 languages that have been enhanced to distinguish between easily confused language pairs (Malay and Indonesian, or Spanish and Portuguese, for example). In using this model, we provide language 'hints,' which increase the prior probability of the ABYZ language labels slightly. The language identification system released by Bergsma *et al.* (2012) is based on compression language models and is designed specifically for use on short inputs. It contains pre-trained models for identifying 261 human languages. We use both language identification models to automatically identify the language of each line of text in the web crawl data.

Qualitatively, we observe that the two language identification systems have different language-specific strengths and weaknesses. The system released by Bergsma *et al.* (2012) is trained on Twitter data and is not always robust to languages written in non-roman alphabets. For example, its Vietnamese model does not recognize the language written in its native script. In contrast, the CLD2 system consistently recognizes Vietnamese. The CLD2 system, however, frequently mistakes Zulu for Swedish because, even when Zulu is suggested as a hint language, Swedish has a much higher prior probability. The Bergsma *et al.* (2012) model consistently recognizes Zulu. Additionally, for some languages in our set, one or both systems don't contain pre-trained models.

We use both sets of pre-trained models to identify the most likely language for each line of text in our web crawls, and we generate a high-precision set of monolingual data consisting of text for which either automatic language identification tag matches the prior newspaper language labels. Table 11.3 in Appendix 11 shows the full amount of data collected for each language as well as the size of each high precision language identification dataset. For most languages, the high precision set contains at least 90% of the full web crawls dataset, suggesting that language labels are correct for most of our web crawls. However, for other languages the high precision set is only a small fraction of the original. This is the case, for example, for Bosnian, Cebuano, Chinese, Kurdish, Serbian, and Uzbek. In the case of Bosnian and Serbian, these languages are closely related and are frequently identified as one other, or as Croatian, by our automatic language identification systems. Uzbek is written using both the Latin and the Cyrillic scripts, however our language identification systems only recognize its Latin script form. For these languages, the relatively small size of the high precision set is a result of errors by the language identification systems. However, in the case of Cebuano,

[10] www.abyznewslinks.com/

[11] https://code.google.com/p/cld2/

Figure 3.2: Word cloud illustrating the amount of monolingual data (web crawls and Wikipedia) that we have gathered and paired with English for 144 languages. Bigger fonts indicate more monolingual data. We have over 200 million words for the following languages, which are not shown in this plot: Russian, Spanish, Farsi, French, Urdu, German, and Italian.

Kurdish and Chinese, the full web crawls datasets contain a large amount of text in other languages, primarily English, and, in generating the high precision set, we eliminate such text. In all of the experiments in this thesis, unless otherwise noted, when we refer to our web crawls datasets, we are referring to the high precision datasets.

Because our data is comprised of news stories, each document also has an associated time stamp, which we use to define a rough document alignment with English news articles. That is, we treat the set of all foreign language news stories published on a particular day as roughly comparable to those written in English on the same day. The degree of comparability between such sets of documents varies greatly. We don't attempt to divide articles published on a given day by topic and infer a finer-grained alignment with English articles on the same topic. We leave this for future work and, instead, use all available data.

In Irvine *et al.* (2014), we describe the American Local News Corpus, which is a parallel web crawl collection effort focused on U.S newspapers. Our methodology for setting up the automatic crawlers and doing deduplication and preprocessing is the same for the two corpora.

3.3.2 Wikipedia

We also use Wikipedia as a source of monolingual data. For all languages, we use Wikipedia's January 2014 data snapshots. To maximize the degree of comparability between our source language Wikipedia pages and English Wikipedia, we only use those pages which have interlingual links with English pages. Unlike our newspaper web crawls, Wikipedia content has fairly reliable language labels. However, for some languages, English content is copied from the English Wikipedia without translation. We use the CLD2 language ID system to identify and remove English content from other languages' Wikipedias. The amount of Wikipedia data that we use for each language is given in Table 11.3 in Appendix 11.

We also use Wikipedia as a source for example *transliterations* in non-roman script languages paired with English. In Irvine *et al.* (2010b), we detailed how we mined transliteration training data from Wikipedia page titles for 150 languages. Wikipedia categorizes articles and maintains lists of all of the pages within each category. In mining transliteration data, we took advantage of a particular set of categories that list people born in a given year. For example, the Wikipedia category page '1961 births' includes links to the 'Barack Obama' and 'Michael J. Fox' pages. We iterated through birth years and the links to pages about people born in each year and then followed interlingual links from each English page about a person, compiling a large list of person names (Wikipedia page titles) in many languages. In Section 4.2.1.3, we use this data to train transliterators and transliterate source language words before comparing their orthographies with English words.

Figure 3.2 illustrates the total amount of monolingual data (web crawls and Wikipedia) that we have for 144 of our 151 languages. We have over 1 billion words of monolingual data for Russian and Spanish and over 200 million for Farsi, Urdu, French, German, and Italian. These languages are not shown in the figure.

3.4 Analysis

In Irvine *et al.* (2013a), we presented a taxonomy of error types related to lexical choice in machine translation as well as two novel techniques for measuring the different types of machine translation errors. In the original paper, we focused on the challenge of adapting machine translation models to new domains of text, and in Chapter 8, we present our analysis of errors in a domain adaptation setting. Here, we measure the number of lexical choice errors of each type that are made in low resource translation. We begin by outlining the error taxonomy and each analysis approach given by Irvine *et al.* (2013a) and then present an analysis of what goes wrong when we do machine translation in low resource conditions.

3.4.1 Approach

3.4.1.1 Error Taxonomy

Errors are categorized into a taxonomy of four error types, S^4: (1) SEEN, (2) SENSE, (3) SCORE, and (4) SEARCH. SEEN errors are a result of some source language phrases being unseen in training, or out-of-vocabulary. SENSE errors are a result of source language phrases being seen in training but not with the correct target language translation. When a source language phrase is observed in training with its correct translation, but that translation is scored lower than another translation alternative during decoding, a SCORE error has been made. Finally, SEARCH errors are a result of pruning translation options during beam search. These four error types account for all translation errors due to *lexical choice* as opposed to word order.

3.4.1.2 Word Alignment Driven Evaluation

The micro analysis, WADE (Word Alignment Driven Evaluation), measures errors at the level of source language words. WADE is based on the fact that we can word align a source language test sentence and its reference translation in the target language. Additionally, the MT decoder naturally produces a word alignment between each input source sentence and its output machine translation. WADE then checks whether the MT output has the same set of target language words aligned to each source language word that we would hope for, given the reference.

In WADE, the unit of analysis is each word alignment between a source language word, f_i, and a reference target language word, e_j. To annotate the aligned pair, $a_{i,j}$, we consider the word(s), H_i, in the output sentence which are aligned (by the decoder) to f_i. If e_j appears in the set H_i, then the alignment $a_{i,j}$ is marked correct. If not, the alignment is categorized with one of the S^4 error types. If the source word f_i does not appear in the phrase table used

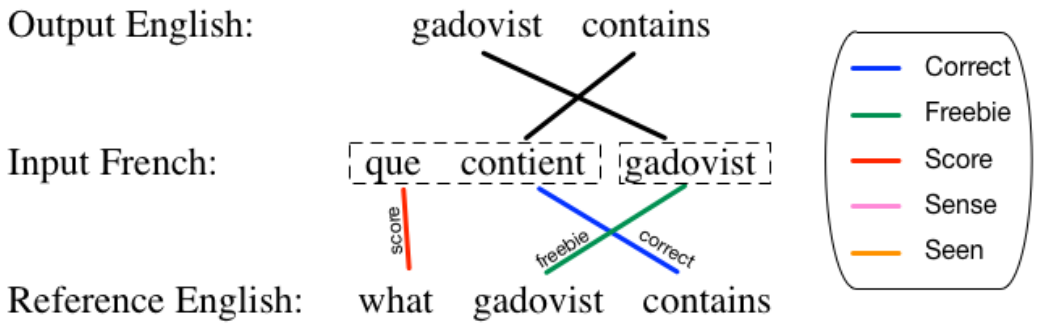


Figure 3.3: Example of WADE visualization. Dashed boxes around the French input mark the phrase spans used by the decoder. In this example, *contient* is translated correctly. The French word *gadovist* does not appear in the phrase table, but because its identity translation is correct, it is marked as a *Freebie*. French *que* is translated as the empty string, but the reference translates it as *what*. Because *what* was a possible translation for *que* in the model but it was not chosen by the decoder, that alignment link is marked as a SCORE error.

for translation, then the alignment is marked as a SEEN error. If f_i does appear in the phrase table, but it is never observed translating as e_j, then the alignment is marked as a SENSE error. If f_i had been observed translating as e_j, but the decoder chose an alternate translation, then the alignment is marked as a SCORE error. For WADE, SCORE errors correspond to errors in phrase table translation scores. Because WADE uses alignment links as its unit of analysis and is agnostic to the word order in the output, it does not capture reordering errors. The results in Irvine *et al.* (2013a) show that search errors are very infrequent, so, following that work, we mark all errors other than SEEN and SENSE as SCORE errors. We make use of one additional category: Freebie. Our MT system copies unseen (OOV) source words into the output, and "freebies" are source words for which this is correct. For related languages written in the same script, like French or Spanish and English, freebies are fairly common. Because WADE's unit of analysis is each *alignment* link between the source text and its reference, it ignores unaligned words in the input source text.

Figure 3.3 shows an example of a WADE-annotated French to English translation. In addition to providing an easy way to visualize and browse the errors in MT output, WADE allows us to aggregate counts over the S^4 error types.

One potential shortcoming of WADE is that it relies on word alignments between our test and reference sets, and alignment errors will result in errors in the error analysis. That is, if a word, f_i, in a sentence in our test set is incorrectly aligned with a word, e_j, in the corresponding reference sentence, WADE will incorrectly indicate an error in the output machine translation if the decoder does not also produce word e_j from f_i. In Appendix 12, we compare WADE analyses based on automatic alignments with those based on manual word alignments which we gather using MTurk. Although the results change with different sets of word alignments, we find that the general trends gleaned from the analyses are consistent.

3.4.1.3 Table Enhancement for Translation Analysis

The macro analysis, TETRA (Table Enhancement for Translation Analysis), artificially supplements a baseline model with different improvements taken from a pseudo-oracle model in order to measure deficiencies in the baseline and potential sources of improvement. This type of analysis is appropriate for the domain adaptation case where a natural comparison is between a baseline OLD-domain model and a pseudo-oracle NEW-domain model. In the low resource case, we approximate a pseudo-oracle with a MT model trained in high resource conditions. We use the high resource model to propose enhancements to the low resource system. This provides a realistic measure of what could be achieved at a corpus level if each error category were targeted for improvement. Here, our experiments are conducted using phrase-based SMT systems, so the translation models (TM) that are enhanced are the phrase table and reordering table. We make three TM enhancements:

- +SEEN enhancement. In order to estimate the effect of SEEN errors, we enhance the TM of the low-resource (LR) model by adding phrase pairs that translate words found only in the high-resource (HR) model, and we measure the BLEU improvement. More precisely, we identify the set of phrase pairs in the HR TM for which the source side contains at least one word that does not appear in the LR training data. These are the phrases responsible for

SEEN errors. We build the system TETRA+seen by adding these phrases to the LR model. When adding these phrases, we add them together with their feature value scores.

- +SENSE enhancement: Analogously, the phrase pairs responsible for SENSE errors are those from the HR model where the source side exists in the LR model, but their English translations do not. We build TETRA+SENSE by adding these phrases to the LR model.

- +SCORE enhancement: To isolate and measure the effect of phrase scores, we consider the phrases that our LR and HR systems have in common. For each phrase pair in the intersection of the two tables, we replace the LR feature scores with the HR feature score. In the case that a phrase pair appears in the LR table and not the HR table, which occurs rarely as a result of word alignment differences between the two models, we include the phrase pair with the LR feature scores. We replace phrase table translation scores and the lexicalized reordering scores in our reordering table separately and then together. We compare the TETRA+translation-scores, TETRA+reordering-scores, and TETRA+all-scores results with the baseline LR model performance.

3.4.2 Experiments

In the experiments presented here, we focus on the following four languages paired with English: Bengali, Hindi, Urdu, and Spanish. Bengali and Hindi paired with English are truly low resource language pairs. We use the small Bengali and Hindi datasets released by Post *et al.* (2012). Training data consists of up to four translations of about 9 thousand Bengali and Hindi sentences. For our WADE learning curve analyses, we simulate typical single-reference training sets by randomly sampling single English translations of each source language training sentence. For the TETRA analyses, we use the full datasets to train our high resource models. Because not a lot of parallel data is available for these languages, the 'high resource' models that we use for the TETRA analyses are also impoverished. That is, we augment low resource models with SEEN, SENSE, and SCORE error corrections estimated from only slightly higher resource models. In all experiments, we use the four-reference tuning and test sets for each language released by Post *et al.* (2012).

We use the 2009 Urdu-English NIST parallel dataset for the Urdu analysis. The training portion of this dataset consists of $88,108$ parallel sentences, or about 1.6 million Urdu and English words.[12] We use the development portion of the NIST dataset for tuning each model and the 883 four-reference test set sentences for the analyses.

For Spanish-English, we use the Europarl v5 parallel corpus (Koehn, 2005). This corpus consists of about 1.7 million parallel sentences, nearly 50 million words of each language. Additionally, we use approximately $2,500$ single-reference parallel sentences each for tuning and testing. The tuning and test sets are newswire articles and are taken from the 2010 WMT shared task (Callison-Burch *et al.*, 2010).[13]

For each of the four language pairs, we randomly sample increasing numbers of sentence pairs and present learning curves based on a WADE analysis. For the TETRA analysis, we use sampled training sets of $1,000$ sentences as low resource (LR) models and the full training sets as high resource (HR) models. We do not attempt to hold the amount of data used to train our HR models constant across languages but, rather, use as much data as possible for each. We use GIZA++ to align each parallel corpus and extract a translation model with a phrase limit of five words. We use two 5-gram English language models in all experiments trained on: (1) the English side of the Europarl corpus, and (2) the English side of the full training corpus for the given language pair.[14] We retune the parameters of each model using batch MIRA (Cherry and Foster, 2012).

3.4.2.1 WADE Analyses

In the cases that we have multiple test set reference translations (Bengali, Hindi, and Urdu), we use the single reference translation that yields the highest BLEU score, given a particular machine translation output.[15] Figure 3.4 gives an example Bengali test set sentence and its translation under two different low resource models: one trained on one thousand parallel sentences and the other on eight thousand parallel sentences. The reference translations that yield the highest BLEU scores are slightly different. Moving from the first model to the second, two SEEN errors are corrected. All of the test-reference alignment links are erroneous under the first machine translation and 50% are under the second.

[12] NIST also released an Urdu-English dictionary, which we do not use in our analyses here.
[13] *news-test2008* plus *news-syscomb2009* for tuning and *newstest2009* for testing.
[14] For Spanish-English, we just use one language model since the training data is Europarl.
[15] In Section 6.4.4 we present a multi-reference version of WADE.

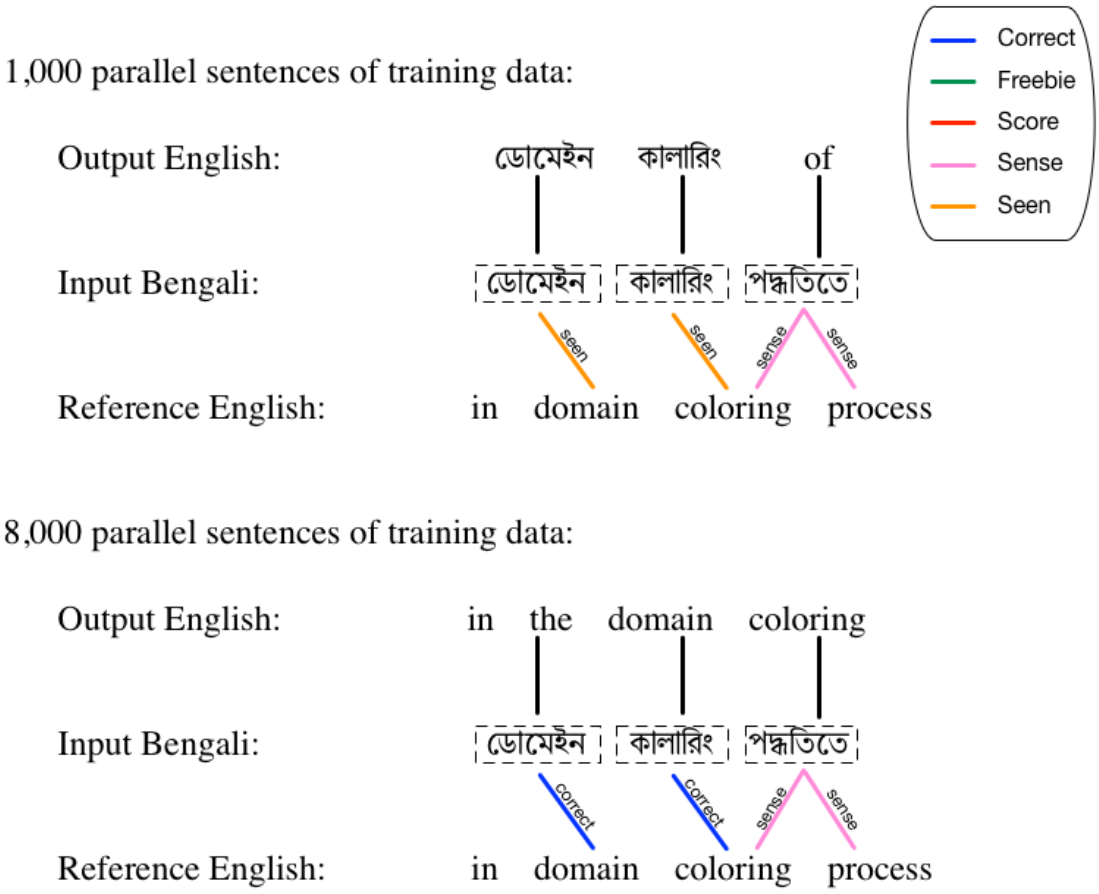
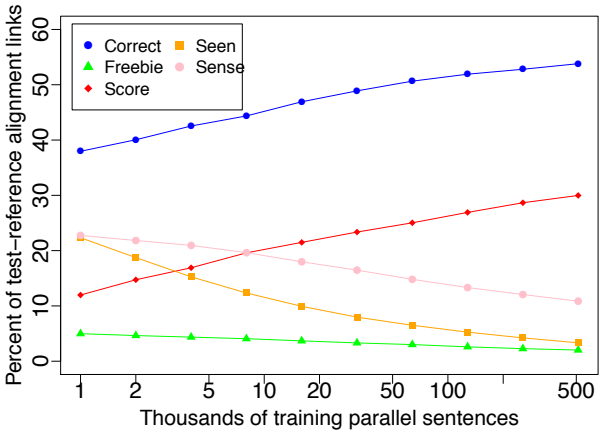


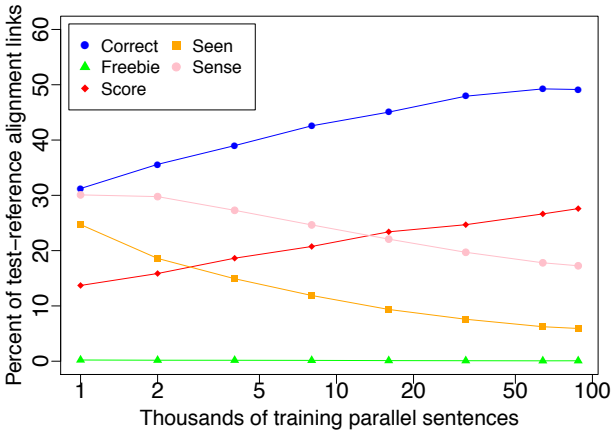
Figure 3.4: WADE analysis on a single Bengali sentence translated using SMT models trained on (1) $1,000$ parallel sentences, and (2) $8,000$ parallel sentences. The reference translations that yield the highest BLEU scores for each machine translation vary slightly. In both cases, there are four alignment links between the Bengali test sentence and the reference translation. Using the model trained on only one thousand sentence pairs, two of the alignment links are marked as SEEN errors and the other two as SENSE errors. Using a model trained on eight thousand sentence pairs, the SEEN errors are corrected, but the SENSE errors remain.

As we discussed in Irvine *et al.* (2013a), WADE allows for easy visualization of machine translation errors at the sentence level as well as aggregate data statistics that describe errors at the corpus level. For example, using the Bengali-English SMT model trained on $1,000$ parallel sentences, only 22% of alignment links in the test set are correct and 78% are incorrect. Moving to the model trained on $8,000$ sentence pairs, 32% are correct and 68% are incorrect.

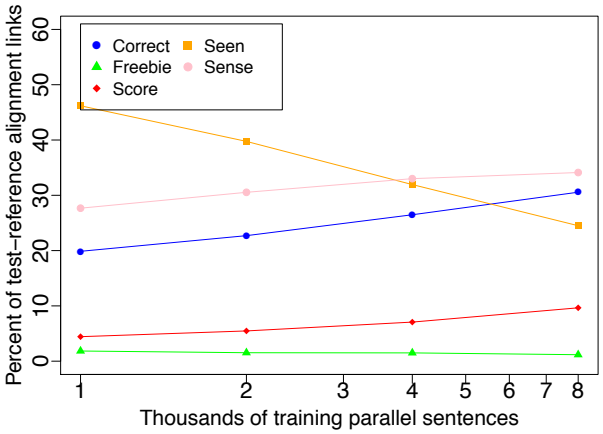
Figure 3.5 shows how the distribution of WADE error types changes with varying amounts of training data for Spanish, Urdu, Bengali, and Hindi translation into English. The learning curves are very similar to one another. With very little training data, SEEN and SENSE errors are the major sources of error. As the amount of training data increases, the numbers of SEEN and SENSE errors decrease. Some of the errors are corrected while others become SCORE errors. SCORE errors indicate that a given translation model has good coverage but translation alternatives are scored in such a way that incorrect target translations are chosen in decoding. For both Spanish and Urdu, SCORE errors are the biggest source of error when models are trained on at least 20 thousand parallel sentences. For Bengali and Hindi, fewer than 10 thousand parallel sentences are available for training, and, under the WADE analysis, SEEN and SENSE errors are the main cause of translation errors when all available training data is used to train the translation models.



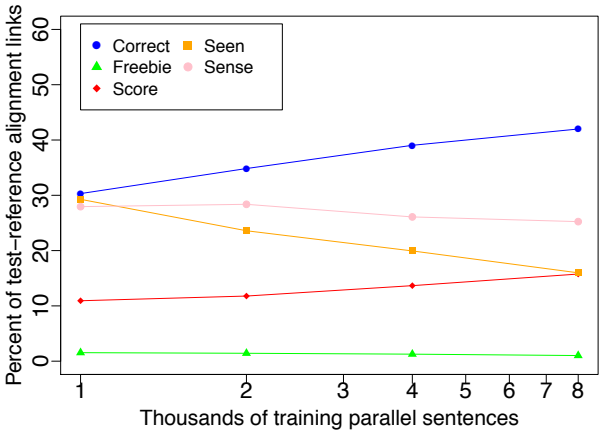
(a) Spanish-English



(b) Urdu-English



(c) Bengali-English



(d) Hindi-English

Figure 3.5: Aggregate WADE analyses on Spanish and Urdu test sets translated using SMT models trained on varying amounts of training data. For both source languages, The number of SEEN and SENSE errors decrease and the number of correct alignments increases with the log of the training data. Although some SEEN and SENSE errors are corrected, many become SCORE errors.

3.4.2.2 TETRA Analyses

We also use a TETRA analysis to answer the question of what goes wrong in translation models trained in low resource conditions. Here, for each of the four source languages, we compare the performance of low resource (LR) models trained on 1 thousand parallel sentences with high resource (HR) models trained on all available training data for the given language pair. We use TETRA to augment each LR model with components from each HR model. Table 3.1 shows the results in terms of BLEU score. Augmenting the LR models with scores from the HR models makes a very small impact. In fact, the HR scores actually hurt performance slightly for three of the four language pairs. In contrast, augmenting the LR models with translations of previously unseen words and phrases yields consistently substantial improvements in BLEU scores, from 2.1 points for Hindi to 6.5 for Spanish. The gains from adding new sense translations are also high and range from 1.8 points for Spanish to 3.7 for Urdu.

The results of our TETRA analysis are consistent with what we found in the WADE analysis: in very low resource conditions, SEEN and SENSE errors are more problematic than SCORE errors. This result is intuitive; in low resource conditions models are estimated over small amounts of parallel data and have low *coverage*. That is, the models don't contain any or all of the possible translations for most source phrases. As the amount of parallel training data increases, estimated translation models have higher coverage but are less precise and more prone to scoring errors.

Language	LR	LR +SEEN	LR +SENSE	LR +SCORE-t	LR +SCORE-r	LR +SCORE-b	HR
Bengali	4.03	6.39 (+2.4)	**6.65 (+ 2.6)**	3.79 (−0.2)	4.03 (+0)	3.72 (−0.3)	12.70
Hindi	6.23	**8.35 (+ 2.1)**	8.11 (+1.9)	6.45 (+0.2)	5.67 (−0.6)	6.08 (−0.2)	15.56
Urdu	7.22	**11.64 (+ 4.4)**	10.90 (+3.7)	7.96 (+0.7)	7.23 (+0)	7.67 (+0.5)	20.62
Spanish	11.35	**17.82 (+6.5)**	13.18 (+1.8)	11.46 (+0.1)	11.25 (−0.1)	11.23 (−0.1)	22.71

Table 3.1: TETRA BLEU score results. For each language, low resource (LR) models are trained on $1,000$ sentence pairs, and high resource (HR) models are trained on the full available datasets (see Section 3.4.2). We use each language's HR model to artificially correct each LR model's SEEN, SENSE, and SCORE errors. We correct phrase table translation SCORE errors (SCORE-t) and reordering SCORE errors (SCORE-r) separately and then together (SCORE-b). For each language, the TETRA augmentation that improves the performance of the LR model the most is highlighted.

3.4.3 Word Alignment Errors

In both the WADE and the TETRA analyses, we are agnostic to the difference between errors that occur as a result of insufficient training data and those that occur as a result of word alignment errors. For example, a SENSE error may occur because a source language phrase truly does not translate as a particular target language phrase in the parallel data. In contrast, SENSE errors may also occur because a source language phrase is incorrectly word aligned, and, as a result, our models don't include the correct target translation in our translation grammars. Disentangling these effects provides insight into potential performance gains from improving word alignments over the small parallel training texts versus moving beyond the parallel training data to attack each error type. One way to disentangle these effects would be to learn a model from manual word alignments over the training data and then compare errors made by the new model with those made by the model based on automatic word alignments. However, manual word alignment is difficult and time-consuming. Instead, we use the word alignment models learned in high resource conditions as a proxy for manual alignments (Callison-Burch *et al.*, 2004). Specifically, we perform additional experiments for Spanish-English and Urdu-English translation comparing the following two word-aligned training datasets:

1. $1,000$ parallel sentences word-aligned using an alignment model estimated over the $1,000$ sentence pairs (LR alignments).

2. $1,000$ parallel sentences word-aligned using an alignment model estimated using a high-resource training dataset (HR alignments).

For Urdu, our high resource word alignment model is estimated over 1.6 million words of training data and for Spanish, it is estimated over 50 million words of training data. Both datasets are described in Section 3.4.2.

Table 3.2 shows the results of our experiments disentangling the effects of the limited coverage of small training sets and poorly estimated word alignment models. BLEU scores go up by about 3 points for Urdu when moving from the low resource alignments to the high resource alignments. This is a considerable improvement in translation quality resulting from improving word alignments alone. For Spanish, the BLEU score improves by about 1.5 points. For both language pairs, most of the gain in translation quality comes from a reduction in SENSE errors. Although some previously SENSE errors become SCORE errors, others are corrected. This is expected: although some source language phrases were observed in training, when they were misaligned, incorrect translations were extracted and the correct translations were not included in the grammars. SENSE errors are particularly pronounced when training data is sparse and source phrases are observed infrequently, reducing the chances that they are aligned with correct translations.

3.4.4 Analysis Conclusion

Our analysis has shown that when we train translation models with only small amounts of parallel data, the major sources of error are SEEN and SENSE errors. That is, many source language words and phrases are not observed at all in training and many others are observed without all of their correct target language translations. In this thesis we improve the performance of translation models learned over small amounts of parallel data by identifying new source word and phrase translations using monolingual corpora. These new translations reduce the number of SEEN and SENSE errors. We further improve performance by adding feature functions estimated over comparable corpora that help models discriminate between good and bad translations, correcting SCORE errors.

	Correct	Freebie	Seen	Sense	Score	BLEU
Urdu						
LR Alignments	31.2	0.2	24.7	30.1	13.8	7.22
HR Alignments	37.7	0.2	23.4	21.8	16.9	10.20
Spanish						
LR Alignments	38.0	5.0	22.3	22.8	12.0	11.35
HR Alignments	40.5	4.9	21.7	17.3	15.5	12.94

Table 3.2: Disentangling the effect of incorrect training data word alignments from the effect of a limited training dataset. The percent of test set and reference translation word alignment links that are correct, freebies, or seen, sense, or score errors are given along with BLEU scores. In both cases, a translation model is estimated over $1,000$ parallel sentences. The low resource (LR) alignment model is estimated using only the $1,000$ parallel sentences, and the high resource (HR) alignment model is estimated using much larger parallel datasets.

Chapter 4

Bilingual Lexicon Induction

Bilingual lexicon induction is the task of inducing word translations from monolingual corpora in two languages. Being about to mine translations from monolingual text is potentially very powerful. In the case of low resource machine translation, we often only have access to a small seed parallel corpus or a small (incomplete) bilingual dictionary as the only bilingual resource available to translate entire texts. Therefore, there are likely to be many unknown (out-of-vocabulary, or OOV) words in the text of interest. In Section 3.4, we showed that unseen words (SEEN errors) are a major source of error in low resource SMT settings and, as we will show in Chapter 8, the same is true in domain adaptation settings. Being able to mine translations for these words from monolingual corpora means that we could produce some translation for every word in our text, achieving perfect model coverage (but not perfect accuracy).

In this chapter, we focus on the task of bilingual lexicon induction. Later, in Chapters 6 and 8, we integrate mined translations into full, end-to-end low resource SMT models and domain adapted SMT models, respectively. We conduct simulated low resource experiments by withholding some of the translations in our bilingual dictionary for evaluation. We use the rest of the dictionary to inform the learning of a translation induction model. This setting mimics what we expect from an actual low resource translation setting: we know how to translate some, but not all, foreign language words.

4.1 Motivating Prior Work

As reviewed in Section 2.2.1, several monolingual distributional similarity metrics, including context, temporal, and orthographic similarity have been proposed as signals that words are translations. Most prior work has used unsupervised methods (like rank combination) to aggregate these types of orthogonal signals (Klementiev and Roth, 2006; Schafer and Yarowsky, 2002). Surprisingly, no past research has employed *supervised* approaches to combine diverse monolingually-derived signals for bilingual lexicon induction. The field of machine learning has shown repeatedly that supervised models dramatically outperform unsupervised models, including for closely related problems like statistical machine translation (Och and Ney, 2002).

For the bilingual lexicon induction task, a supervised approach is natural, particularly because computing contextual similarity typically requires a seed bilingual dictionary (Rapp, 1995), and that same dictionary may be used for estimating the parameters of a model to combine monolingual signals. Alternatively, in a low resource machine translation (MT) setting, it is reasonable to assume a small amount of parallel data from which a bilingual dictionary can be extracted for supervision. In this setting, bilingual lexicon induction is critical for translating source words which do not appear in the parallel data or dictionary.

Previous work in bilingual lexicon induction only reports results on inducing translations for the most frequent source language words (often only nouns), completely avoiding any scalability or data sparsity issues. Because those word counts are not sparse, that task is much easier than inducing translations for a random set of words, as Section 4.3 shows. In end-to-end SMT, unknown words tend to be relatively infrequent. This means that it is unclear whether previous bilingual lexicon induction results would improve SMT quality for low resource languages. In Section 4.3, we present experimental results on a wide variety of languages, for which a wide variety of monolingual corpora and seed bilingual dictionaries are available. We also present bilingual lexicon induction results in terms of monolingual word frequencies in order to understand the effects of data sparseness.

4.2 Using Monolingual Data to Predict Translations

We frame bilingual lexicon induction as a binary classification problem; for a pair of source and target language words, we predict whether the two are translations of one another or not. For a given source language word, we score all target language candidates separately and then rerank them. We use a variety of signals derived from source and target monolingual corpora as features and use supervision to estimate the strength of each. Our diverse set of signals, which serve as features in our classification framework, include contextual, temporal, topical, orthographic, frequency, and burstiness similarity. We presented our method for doing minimally supervised bilingual lexicon induction in Irvine and Callison-Burch (2013b), and this chapter extends upon that work.

We describe the diverse set of monolingual signals in Section 4.2.1, several of which we first proposed in Klementiev *et al.* (2012). We report the performance of individual signals alone and a baseline method for combining them in Section 4.2.3. Then, in Section 4.2.4, we use a supervised method for learning to combine the signals to predict word translations.

4.2.1 Monolingual Signals of Translation Equivalence

4.2.1.1 Contextual Similarity

We use the vector space approach of Rapp (1999) to compute similarity between word in the source and target languages. More formally, assume that $(s_1, s_2, \ldots s_N)$ and $(t_1, t_2, \ldots t_M)$ are (arbitrarily indexed) source and target vocabularies, respectively. A source word f is represented with an N-dimensional vector and a target word e is represented with an M-dimensional vector (see Figure 2.3). The component values of the vector representing a word correspond to how often each of the words in that vocabulary appear within a two word window on either side of the given word. These counts are collected using monolingual corpora. After the values have been computed, a contextual vector f is projected onto the English vector space using translations in a given bilingual dictionary to map the component values into their appropriate English vector positions. This sparse projected vector is compared to the vectors representing all English words, e. Each word pair is assigned a contextual similarity score $c(f, e)$ based on the similarity between e and the projection of f.

Various means of computing the component values and vector similarity measures have been proposed in literature (e.g. Fung and Yee (1998); Rapp (1999)). Following Fung and Yee (1998), we compute the value of the k-th component of f's contextual vector, f_k, as follows:

$$f_k = n_{f,k} * (log(n/n_k) + 1)$$

where $n_{f,k}$ and n_k are the number of times s_k appears in the context of f and in the entire corpus, and n is the maximum number of occurrences of any word in the data. Intuitively, the more frequently s_k appears with f_i and the less common it is in the corpus in general, the higher its component value. After projecting each component of the source language contextual vectors into the English vector space, we are left with M-dimensional source word contextual vectors, F, and target word contextual vectors, E, for all words in the vocabulary of each language. We use cosine similarity to measure the similarity between each pair of contextual vectors:

$$sim_{context}(F, E) = \frac{F \cdot E}{||F||||E||}$$

Table 4.1 shows example ranked lists using contextual similarity to rank English words for several Spanish words. For example, contextual similarity ranks the English words *reached*, *enjoyed*, and *contained* highly as candidate translations of Spanish *alcanzaron*. These incorrect English words tend to appear in similar contexts as the correct English translation, *reached*. In Appendix 13.1, we present results using a variety of both probabilistic and non-probabilistic dictionaries to project contextual vectors.

4.2.1.2 Temporal Similarity

Some of our monolingual corpora have associated time stamps. In particular, each document in our web crawls of online news websites has an associated publication date (see Section 3.3). We gather temporal signatures for each source and target language unigram from our time-stamped web crawl data in order to measure temporal similarity (Alfonseca *et al.*, 2009; Klementiev and Roth, 2006; Schafer and Yarowsky, 2002). The intuition is that news stories

alcanzaron	sanitario	desarrollos	volcánica	montana
reached	exil	advances	**volcanic**	arendt
enjoyed	rhombohedral	**developments**	eruptive	**montana**
contained	apt	changes	coney	glasse
contains	immune	placing	rhonde	teter
saw	circulatory	innovations	bleaker	waddingham
includes	nervous	use	staten	daryl
included	endocrine	changes	robben	callowhill
hit	coordinate	making	ostrov	richings
achieved	ucsd	addition	ellesmere	beswick
estates	windowing	allowing	gilligan	holgersson

Table 4.1: Example of ranked word translations using contextual similarity. The correct English translations, when found, are bolded. English words are ordered by their contextual similarity scores with the given Spanish word.

in different languages will tend to discuss the same world events on the same day and, correspondingly, we expect that source and target language words which are translations of one another will appear with similar frequencies over time in monolingual data. For instance, if the English word *tsunami* is used frequently during a particular time span, the Spanish translation *maremoto* is likely to also be used frequently during that time. Figure 4.1 illustrates how the temporal distribution of Spanish *terremoto* is more similar to its English translation *earthquake* than to other English words. *Microsoft*, one of the non-translations, like *earthquake*, is very bursty (formal definition given in Section 4.2.1.6). *Strength*, another non-translation, in contrast, appears with fairly consistent frequency over time. The temporal histograms for *terremoto* and *earthquake* both show significant peaks in the middle of the series, which correspond to the major earthquake that occurred in Haiti in January of 2010. Although the two words have reasonably well matched temporal signature, there are some differences. For example, there is an earthquake event near the end of the series that is covered in Spanish news but not as much in English news.

We calculate the temporal similarity between a pair of words, $sim_{temp}(F, E)$ using the method defined by Klementiev and Roth (2006). We generate a temporal signature for each word by sorting the set of (time-stamped) documents in the monolingual corpus into a sequence of equally sized temporal bins and then counting the number of word occurrences in each bin. Our English web crawl data is essentially limitless, so we restrict the English data that we use in a particular foreign language experiment to be no more than three times the size of our source language web crawled data, and only include news articles from those dates for which we also have source language articles. In our experiments, we set the temporal bin size to 3 days, so the size of temporal signatures is equal to the number of days spanned by our corpus divided by three. We normalize the temporal signature of each word by dividing all of the counts by the total count and, again, we use cosine similarity to compare the normalized temporal signatures for a pair of words:

$$sim_{temp}(F, E) = \frac{F \cdot E}{||F||||E||},$$

where F and E are source and target language word temporal signatures, respectively.

Table 4.2 shows example ranked lists using temporal similarity to rank English words for several Spanish words. For example, *ash* and *spewed*, as well as the Icelandic volcano *eyjafjallajokull*, are all temporally similar to the Spanish word *volcánico*. Since volcanic eruptions are generally talked about in newspapers all around the world when they occur, it is not surprising that this signal is able to score several related words highly. In Appendix 13.2, we compare the performance of using raw temporal signatures and using the Discrete Fourier Transform of those signatures.

4.2.1.3 Orthographic Similarity

We measure orthographic similarity between a pair of words as the normalized[1] edit distance between the two words:

$$sim_{orth}(f, e) = \frac{ed(f, e)}{\frac{|e||f|}{2}}$$

[1]Normalized by the average of the lengths of the two words

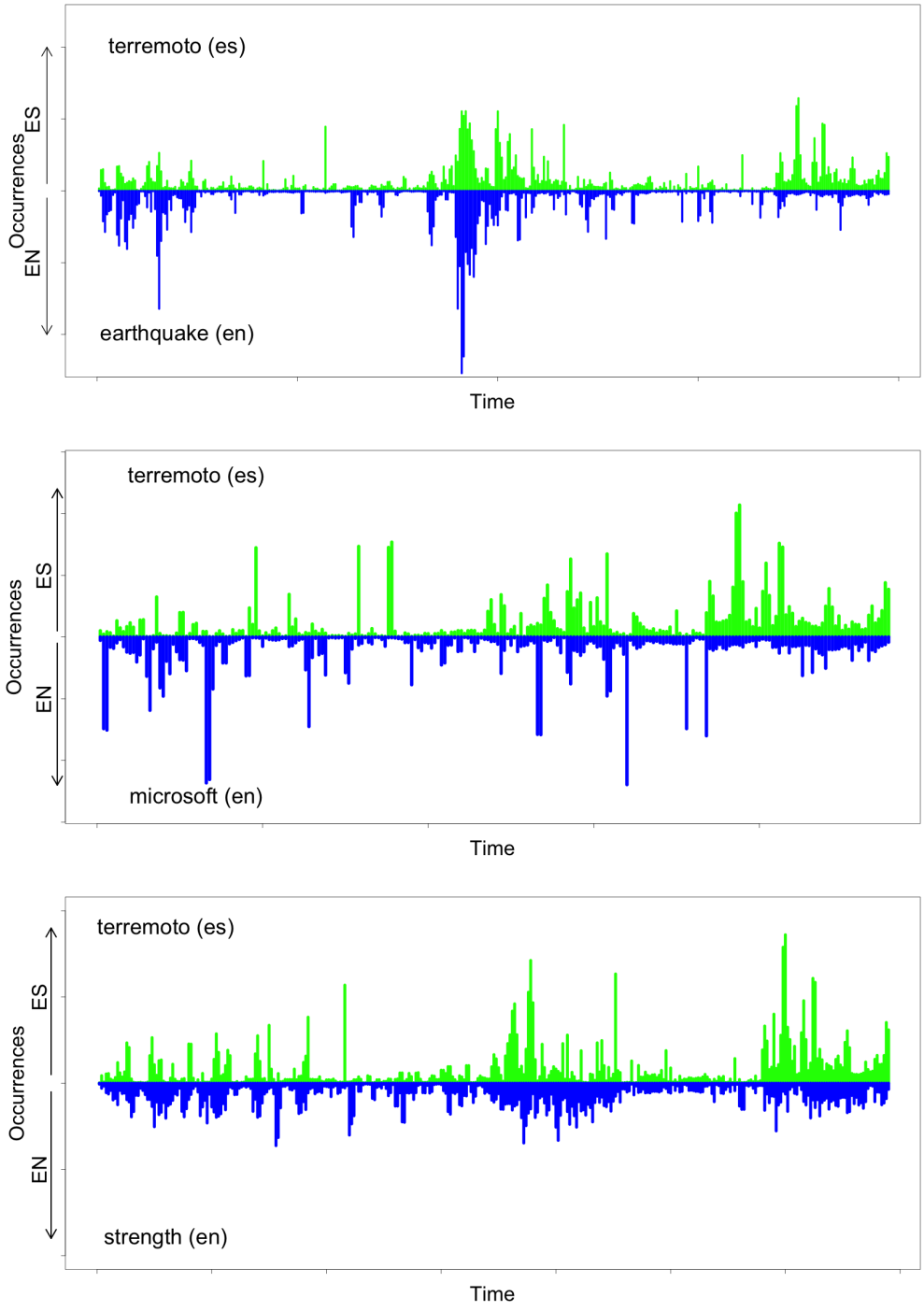


Figure 4.1: Temporal histograms of the Spanish word *terremoto* paired with three English candidate translations: the correct translation *earthquake* and the incorrect candidates *microsoft* and *strength*. The temporal histograms are collected from monolingual texts spanning several years and show the number of occurrences of each word (on the y-axes) across time. While the correct translation has a good temporal match (sim_{temp}(terremoto, earthquake) $= 2 \cdot 10^{-4}$), the non-translations are less temporally similar (sim_{temp}(terremoto, microsoft) $= 2 \cdot 10^{-5}$, sim_{temp}(terremoto, strength) $= 3 \cdot 10^{-5}$). In all examples, only dimensions (dates) which are non-zero valued for both signatures are shown, which results in the signature for *terremoto* appearing somewhat different across the three comparisons.

alcanzaron	sanitario	desarrollos	volcánica	montana
travel	snowpocalypse	occupied	wawel	dzv
road	airport	aer	**volcanic**	spatz
news	dioxide	madoff	ash	centimes
services	steinmeier	declaration	spewed	kleve
arts	gobbling	ponzi	eyjafjallajokull	reallocate
word	investigating	affects	otunbajewa	frostrup
special	convicted	suspected	eruption	roze
chief	spy	fed	cloud	minc
top	offices	combat	rubell	bicyclists
inspired	bond	arrested	dormancy	lgbt

Table 4.2: Example of ranked word translations using temporal similarity. The correct English translations, when found, are bolded. English words are ordered by their temporal similarity scores with the given Spanish word.

Russian→English
f o t → *f a u t*
c y → *t s y*
w u k → *s c h u k*
Greek→English
$o\ \chi\ ´\alpha \rightarrow o\ c\ h\ a$
$\gamma\ \epsilon\ \rho \rightarrow g\ e\ r$
$\alpha\ \lambda\ \mu \rightarrow a\ l\ l\ m$

Table 4.3: Examples of Russian to English and Greek to English transliteration rules extracted from pairs of person names.

where *ed* is the standard Levenshtein edit distance between the two strings. For word classes that tend to be transliterated, rather than translated, (e.g. person and place names, and etymologically related words), we expect the edit distances between English words and their translations to be small. Prior work has learned mappings between character sets (e.g. Snyder *et al.* (2010); Yamada and Knight (1999)). Berg-Kirkpatrick and Klein (2011) use decipherment techniques to learn correspondences between the alphabets of two languages given two lexicons containing unmatched cognates.

For non-Roman script languages, we transliterate words into the Roman script before measuring orthographic similarity, following our prior work in Irvine *et al.* (2010b). We treat transliteration as a monotone character translation task and train models on the mined pairs of person names in foreign, non-Roman script languages and English. Because transliteration is strictly a monotone operation, we do not allow reordering in our models. Additionally, unlike in machine translation, our translation and language models can support very large n-gram sizes because the number of characters in a given script is small compared to word vocabularies; we use phrase length limits of 10 when extracting translation grammars and in estimating language models. We use a character-based language model trained on the complete list of English names. Table 4.3 shows some example rules that we learn for transliterating Russian and Greek into the Roman script.

In Irvine *et al.* (2010b), we provide a detailed evaluation of our transliteration technique. For purposes of bilingual lexicon induction, we use the top-1 transliteration to compute edit distance.

Table 4.4 shows example ranked lists using orthographic similarity to rank English words for several Spanish words. For those Spanish words that have English cognates, such as *sanitario* and *volcánica*, the orthographic signal ranks correct translations highly.

4.2.1.4 Topic Similarity

Words and their translations are likely to appear in articles written about the same topic in two languages. Thus, topic or category information associated with monolingual data can also be used to indicate similarity between a word and its

alcanzaron	sanitario	desarrollos	volcánica	montana
alcantara	**sanitary**	ferroalloy	**volcanic**	**montana**
albanian	sanitation	barrosos	volcanism	fontana
lazzaroni	unitario	destroyers	voltaic	montane
lanaro	sanitarium	mccarroll	vacancy	mentana
aleandro	sanitation	disallows	konica	montagna
lazaros	sagittario	disallow	dominica	montanha
canaro	sanitarias	scrolls	veronica	montan
alianza	kantaro	payrolls	monica	montano
lazaro	sanitorium	carroll	volcano	montani
catanzaro	santoro	steamrolls	vratnica	montand

Table 4.4: Example of ranked word translations using orthographic similarity. The correct English translations, when found, are bolded. English words are ordered by their orthographic similarity scores with the given Spanish word.

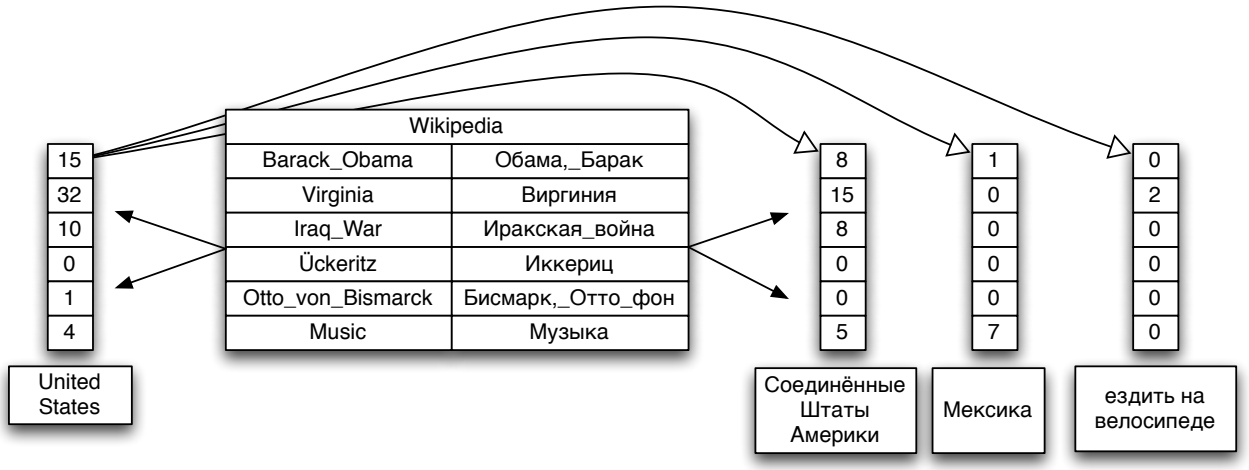


Figure 4.2: Illustration of how we compute the topical similarity between *United States* and three Russian candidate phrase translations. We first collect the topical signatures for each phrase (e.g. *United States* appears in the page about *Barack Obama* 15 times and in the page about *Virginia* 32 times.) based on the interlingually linked pages. We can then directly compare each pair of topical signatures.

candidate translation. In order to score a pair of words, we collect their topic signatures by counting their occurrences in each topic, normalize the signatures, and then comparing the resulting vectors. We, again, use the cosine similarity measure:

$$sim_{topic}(F, E) = \frac{F \cdot E}{||F||||E||},$$

where F and E are normalized source and target language word topical signatures, respectively.

In our experiments, we use interlingual links between Wikipedia articles to estimate topic similarity. We treat each linked article pair as a topic and collect counts for each word across all articles in its corresponding language. Thus, the dimensionality of a word's topic signature corresponds to the number of interlingually linked article pairs, and each value corresponds to the number of times the word appears in the given article. For each foreign language, the number of Wikipedia articles linked to English pages is given in Table 11.3 in Appendix 11. The number of linked articles range from 84 (Kashmiri) to over 500 thousand (French). Figure 4.2 illustrates this signal. Our Wikipedia-based topic similarity signal is similar in spirit to polylingual topic models (Mimno *et al.*, 2009).

Table 4.5 shows example ranked lists using topic similarity to rank English words for several Spanish words. Using topic similarity, *montana*, *miley*, and *hannah* are ranked highly as candidate translations of the Spanish word *montana*. The TV character Hannah Montana is played by actress Miley Cyrus, so the topic similarity between these words makes sense.

alcanzaron	sanitario	desarrollos	volcánica	montana
reached	health	**developments**	**volcanic**	**montana**
began	transcultural	developed	eruptions	miley
led	medical	development	volcanism	hannah
however	sanitation	used	lava	beartooth
early	patient	using	plumes	cyrus
including	deliverables	modern	eruption	crazier
took	pharmaceutical	based	volcano	bozeman
remained	sewerage	important	volcanoes	chelsom
several	healthcare	history	breakouts	absaroka
continued	care	different	volcanically	baucus

Table 4.5: Example of ranked word translations using topic similarity. The correct English translations, when found, are bolded. English words are ordered by their topic similarity scores with the given Spanish word.

4.2.1.5 Frequency Similarity

Words that are translations of one another are likely to have similar relative frequencies in monolingual corpora. We measure the frequency similarity of two words, sim_{freq}, as the absolute value of the difference between the log of their relative corpus frequencies, or:

$$sim_{freq}(e, f) = |log(\frac{freq(e)}{\sum_i freq(e_i)}) - log(\frac{freq(f)}{\sum_i freq(f_i)})|$$

4.2.1.6 Burstiness Similarity

Burstiness is a measure of how peaked a word's usage is over a particular corpus of documents (Pierrehumbert, 2012). Bursty words are topical words that tend to appear when some topic is discussed in a document. For example, *earthquake* and *election* are considered bursty. In contrast, non-bursty words are those that appear more consistently throughout documents discussing different topics, *use* and *they*, for example. Church and Gale (1995, 1999) provide an overview of several ways to measure burstiness empirically. Following Schafer and Yarowsky (2002), we measure the burstiness of a given word in two ways. The first is based on Inverse Document Frequency (IDF):

$$IDF_w = -log\frac{df_w}{|D|},$$

where df_w is the number of documents that w appears in, and $|D|$ is the total number of documents in the collection. The second burstiness measure, similar to that defined by Church and Gale (1995), is the average frequency of w divided by the percent of documents in which w appears. We make one modification to the definition provided by Church and Gale (1995) and use relative frequencies rather than absolute frequencies to account for varying document lengths.

$$B_w = \frac{\sum_{d_i \in D} rf_{w_{d_i}}}{df_w},$$

where, as before, df_w is the number of documents in which w appears and $rf_{w_{d_i}}$ is the relative frequency of w in document d_i. Relative frequencies are raw frequencies normalized by document length. Table 4.6 shows examples of high and low ranked bursty words under each measure for two different constant word frequencies. The examples show that both measures of burstiness yield rankings that are consistent with our intuitions, yet they provide different results.

We compare both the IDF and the B scores for pairs of words using ratios:

$$sim_{IDF}(e, f) = min[\frac{IDF_e}{IDF_f}, \frac{IDF_f}{IDF_e}]$$

$$sim_{burst}(e, f) = min[\frac{B_e}{B_f}, \frac{B_f}{B_e}]$$

Frequency, f, and number of words, n	IDF		Burstiness	
	Top-5	Bottom-5	Top-5	Bottom-5
$f = 50, n = 802$	kratsa tebet kagome khaldūn psittacosaurus	contemporaneously unrecognizable categorizing modern-style crazed	straubing-bogen tebet cloppenburg autosan gøta	wavering busing unconvinced redesigning oftentimes
$f = 100, n = 303$	subarticle trackmania lyrebird gârbea biecz	call-ups workable purports outnumber unmatched	penedès lyrebird azarbaijan padstow trackmania	demoralized misgivings precluded workable forestall

Table 4.6: Examples of highest and lowest ranked English words according to two measures of burstiness. Empirical estimates were taken from a subset of English Wikipedia data.

4.2.1.7 Additional Signals

In addition to the basic features listed above, we perform some experiments to test the usefulness of additional signals of translation equivalence. Two such signals are word prefix contextual similarity and word suffix contextual similarity. Prefix contextual similarity is calculated in the same way as the contextual similarity score, but we use source and target word stems, or word prefixes up to five characters long, instead of full words. That is, the word prefix contextual similarity score for the word pair (*blanc*o, *white*) is the same as that of (*blanc*a, *white*). In this particular example, we collect only a single contextual vector for *blanc*{*o,a*}. In Spanish, this translation of the English word *white* appears with either a masculine or feminine ending, depending on what it modifies. By summing the distributional counts of *blanco* and *blanca*, we expect a contextual vector that is more similar to English *white* than either alone. We measure the similarity of a pair of prefixal contextual vectors using cosine similarity, as before.

Suffix contextual similarity measure is similar to the word stem measure, except instead of using word prefixes, it uses word *suffixes* of up to five characters long. For example, the word stem contextual similarity score of the word pair (impo*sible*, pos*sible*) is the same as that of (po*sible*, impose*sible*). With this signal, we expect to sum over alternate word prefixes in the same way that the word stem signal sums over alternate word suffixes. Again, the similarity between a pair of suffixal contextual vectors is measured using cosine similarity.

In addition to prefix and suffix contextual similarity, we also estimate prefix and suffix topic and temporal similarity. We also use an indicator feature which is positive if the source and target words are the same string. Of course, this indicator is most useful for languages written in the same script. Finally, we add a final feature indicating the target translation's monolingual frequency, which serves as a sort of prior probability that the target word is of interest at all. Specifically, we define this feature as the inverse of the log of the target word's frequency.

Although we have limited our experiments to this set of varied signals of translation equivalence, our basic framework is easily extendible.

4.2.2 Orthogonality of Signals

In this section we seek to answer two questions about how the signals presented in Section 4.2.1 interact. First, intuitively, the signals presented seem orthogonal. That is, they provide very different types of information about how words are used in language, and we hypothesize that the lists of ranked candidate translations under each signal are uncorrelated with the exception (and hope!) that correct translation pairs rank relatively high according to all or most of the signals. In our first set of experiments, we measure their orthogonality empirically. Second, we hypothesize that some signals tend to rank translation candidates more accurately than others. For example, we would expect that the frequency signal is a weaker predictor than, for example, orthographic similarity, particularly for closely related language pairs. In our second set of experiments, we compare the accuracies of each signal and include analyses by language and by part-of-speech.

In order to empirically measure orthogonality of our signals, we measure pairwise Spearman rank-order correlation coefficients. Specifically, we first use each signal separately to rank all translation candidates. Then, we measure the correlation between all pairs of ranked lists using the Spearman coefficient. A correlation coefficient of 1.0 indicates

	crawls-cont	wiki-cont	temporal	orth.	topic	freq.	burst.
wiki-cont	-0.15						
temporal	-0.14	-0.19					
orth.	-0.28	-0.31	-0.28				
topic	-0.15	-0.14	-0.13	-0.30			
freq.	0.01	0.13	0.02	-0.18	0.13		
burst.	-0.10	0.06	-0.07	0.06	0.11	0.28	
idf	0.06	0.10	-0.12	-0.01	0.00	0.49	0.14

Table 4.7: Measure of the correlation (orthogonality) between signals. For each of 24 languages, we randomly select $1,000$ source language words and compute the Spearman rank correlation coefficient across pairwise ranked lists of translation candidates generated by each of eight signals of translation equivalence. We average coefficients within each language. The results here show the mean of the correlation coefficient between all pairs of signals across the 24 languages.

perfect positive correlation, -1.0 indicates perfect negative correlation, and coefficients close to zero indicate that our signals do not correlate.

For each of 24 languages,[2] we randomly select $1,000$ source language words and use each of our eight basic translation signals to rank all candidate English translations. For each source language word and each pair of signals, we measure the Spearman correlation coefficient. We average the pairwise results across the $1,000$ source words and then average across languages.

Table 4.7 shows the results. The first thing to note is that the highest average correlation coefficient is between the frequency and the inverse-document frequency (IDF) signals (0.49). This makes sense because IDF is based on word frequency. The second highest value corresponds to a negative correlation (-0.31) between orthographic similarity and wikipedia contextual similarity. These features are based on entirely different information, and we would not expect them to have a positive correlation. The fact that they are negatively correlated is surprising, but confirms our intuition that the signals provide orthogonal information.

The question of whether different signals tend to better predict the translations of different source words or whether one or two signals always dominate remains. Relatedly, it is possible that some signals are better able to predict the translations of certain *classes* of words that others. For example, it may be true that the orthographic and topic signals are better at predicting the translations of named entities than, for example, the contextual signals, which may better predict the translations of closed class words.

First, we ask how frequently each signal ranks the correct translation higher than any other signal. That is, we ask how often each signal is a better predictor of how to translate a given word than all other signals. We use the same set of randomly selected $1,000$ source language words used to estimate rank correlations. For each, we identify the rank of the *correct* English translation under each of the eight basic signals. We then measure how often each signal ranks the correct translation higher than the other signals. As elsewhere in this chapter, we use the Mechanical Turk (see Section 3.2.2) dictionaries as gold standards.

Table 4.8 shows the results. Although each signal is the most informative for at least some source language words in all 24 languages, the following three dominate most often: wikipedia contextual similarity, orthographic similarity, and topic similarity. Given this result, we ask a related question: are some signals particularly informative for certain classes of words? In order to begin to answer this question, we label each source word with the most frequent part-of-speech (POS) tag for its English translation using the tagger in the Natural Language Toolkit (NLTK).[3] We use the pre-trained sequential backoff tagger released with NLTK, which tags sequences of English words with tags from the Penn Treebank (Marcus *et al.*, 1993). However, we do not tag word sequences but rather individual English words in our test set. We use information from English because POS taggers are not readily accessible for many of our languages of interest.

Given English POS tags projected onto each of our source language words, we perform a similar analysis as before, but we now group source language words by POS tag. Then, for each language and all words in each POS category, we count how often each signal ranks a correct translation higher than all other signals. Table 4.9 shows the results. For clarity, we collapse some POS classes. For example, we mark both noun and plural nouns as simply 'Noun.'

[2] We use the same set of 24 languages that we experiment with elsewhere in this chapter. The languages are listed in Table 4.11.

[3] http://www.nltk.org/

Language	crawls-cont	wiki-cont	temporal	orth.	topic	freq.	burst.	idf
Azeri	3.6	41.0	3.6	11.0	30.3	5.9	4.2	0.4
Bulgarian	5.1	27.0	3.1	17.0	42.2	4.3	0.6	0.8
Bengali	8.7	26.7	0.9	15.4	40.4	4.5	2.3	1.2
Bosnian	8.8	41.2	4.2	16.5	21.8	4.7	2.5	0.4
Cebuano	12.7	22.1	7.3	20.6	25.7	4.6	6.4	0.5
Welsh	11.0	55.6	3.2	9.6	11.1	8.0	1.2	0.4
Gujarati	9.4	33.9	5.3	8.6	31.8	4.3	3.9	2.9
Hindi	4.5	25.5	2.0	10.6	46.7	4.9	2.8	2.9
Hungarian	4.6	36.1	0.0	10.1	25.7	12.5	5.4	5.7
Indonesian	12.3	54.9	4.3	10.8	6.4	7.9	0.5	2.8
Latvian	5.4	41.6	4.8	18.6	23.1	5.0	1.3	0.3
Nepali	11.2	32.0	6.4	12.5	27.6	5.1	4.2	0.8
Romanian	5.7	39.3	1.5	35.0	9.6	5.4	2.7	0.8
Slovak	4.8	42.1	4.2	17.5	22.8	4.3	3.3	1.0
Somali	8.7	28.3	3.4	11.1	18.1	17.4	12.5	0.5
Albanian	7.2	47.8	3.1	21.9	11.0	6.0	3.0	0.1
Serbian	3.8	27.4	1.6	17.5	42.8	4.5	1.6	0.7
Swedish	4.3	45.0	2.1	22.3	10.7	11.1	2.5	2.1
Tamil	7.7	25.2	1.8	4.2	53.7	5.1	1.6	0.8
Telugu	6.6	29.4	5.8	10.2	39.9	3.1	3.4	1.6
Turkish	6.8	43.4	8.7	9.8	15.2	11.4	2.5	2.1
Ukrainian	7.2	35.1	4.0	24.0	17.0	6.9	3.6	2.2
Uzbek	7.4	6.6	0.5	20.1	41.0	15.1	7.4	1.9
Vietnamese	11.0	16.6	9.7	7.7	21.0	16.6	3.3	14.1
Average	7.4	34.3	3.8	15.1	26.5	7.4	3.4	2.0

Table 4.8: Percent of $1,000$ words for which each translation signal ranks a correct translation highest, across 24 languages.

POS Class	% Words	crawls-cont	wiki-cont	temporal	orth.	topic	freq.	burst.	idf
Verb	10.9	8.9	34.0	4.6	7.3	31.1	9.1	2.9	2.1
Noun	64.8	7.0	36.7	3.5	17.4	23.7	7.0	2.9	1.9
Adverb	3.9	10.5	35.3	6.6	5.1	29.0	7.3	3.5	2.6
Adjective	13.3	6.2	34.4	3.1	19.0	27.3	5.5	3.1	1.4
Closed	7.1	9.4	28.4	5.3	6.6	36.8	5.4	7.0	1.1
Average		7.4	34.3	3.8	15.1	26.5	7.4	3.4	2.0

Table 4.9: Analysis of Signals by Part-of-Speech tag.

Because there are so few word types, we also collapse all closed class categories, including conjunctions, determiners, and prepositions into a single 'Closed' category. In comparison with Table 4.8, results are grouped by POS tag instead of by language. The final row is identical to that in Table 4.8. Because most (65%) words are nouns, the summary statistics are dominated by them.

The results in Table 4.9 are very consistent across word classes with one notable exception. The orthographic feature makes very good translation predictions for nouns and adjectives but not for the other word classes. This makes sense; we would expect orthographic similarity to be informative for borrowed and transliterated words, which tend to be nouns. The overall consistency suggests that there is likely little to gain from training word class-specific models for making translation predictions. In Section 4.2.3, we define a baseline method for combining the orthogonal features to make a single translation prediction, and in Section 4.2.4 we *learn* models for combining features.

4.2.3 Individual Monolingual Signals

The features that we use to identify word translations are based on the signals described above, and similarity scores are estimated over two source of comparable corpora, web crawls and Wikipedia (see Section 3.3 for a detailed description of each dataset):

1. Web Crawls Contextual Similarity
2. Web Crawls Temporal Similarity
3. Orthographic Similarity
4. Wikipedia Contextual Similarity
5. Wikipedia Topic Similarity
6. Wikipedia Frequency Similarity
7. Wikipedia IDF Similarity
8. Wikipedia Burstiness Similarity
9. Web Crawls Prefix Contextual Similarity
10. Web Crawls Prefix Temporal Similarity
11. Web Crawls Suffix Contextual Similarity
12. Web Crawls Suffix Temporal Similarity
13. Wikipedia Prefix Contextual Similarity
14. Wikipedia Prefix Topical Similarity
15. Wikipedia Suffix Contextual Similarity
16. Wikipedia Suffix Topical Similarity
17. String Identity

18. Inverse Log of Target Wikipedia Frequency

Table 4.10 shows examples of Romanian words paired with several English words, both correct and incorrect, and scored with all 18 features.

It's intuitive that combining these orthogonal measures would improve performance on bilingual lexicon induction, and Schafer (2006) showed that to be true. We define our baseline for combining our set of monolingual signals, H, to be the mean reciprocal rank (MRR) across all features:

$$MRR_e = \frac{\sum_{h \in H} \frac{1}{r_h(e)}}{|H|}$$

where $r_h(e)$ is the rank of English word e under the monolingual similarity measure h. This unsupervised approach to rank aggregation assumes no prior knowledge of which signals are likely to be the most informative.

We measure performance using accuracy in the top-k ranked translations. We define top-k accuracy over some set of ranked lists L as follows:

$$acc_k = \frac{\sum_{l \in L} I_{lk}}{|L|}$$

where I_{lk} is an indicator function that is 1 if and only if a correct item is included in the top-k elements of list l. That is, top-k accuracy is the proportion of ranked lists in a set of ranked lists for which a correct item is included in the highest k ranked elements.

Figure 4.3 shows the performance of each of the monolingual similarity measures alone as well as the baseline MRR method for combining them. Each box-and-whisker plot shows the top-10 accuracy range, quartiles, and median across a set of 24 diverse languages (listed in Figure 4.11). The Wikipedia topic and context features using whole words and word prefixes are the highest performing single features. More importantly, Figure 4.3 shows that even using the simple MRR method of combining signals is more effective than using any single feature. This result motivates our approach to using supervision to learn how to optimally combine these orthogonal signals and output a more accurate ranking.

4.2.4 Learning to combine orthogonal monolingual signals

We use a *supervised* approach to combining the monolingual signals enumerated above. For each language, we choose up to $10,000$ source language words among those that occur in each of our comparable corpora (web crawls and wikipedia) at least ten times and that have at least one translation in our gold standard dictionaries. In this chapter, in order to keep our experiments consistent across languages, we use the Mechanical Turk (see Section 3.2.2) dictionaries as gold standards. Because some monolingual datasets and some dictionaries are small, the source word samples are smaller than $10,000$ for some languages. For example, although our MTurk dictionary contains translations for $9,977$ Gujarati words, only $4,442$ of those words appear at least ten times in both of our monolingual corpora. We randomly divide the source language words into three equally sized sets for training, development, and testing.

We learn binary classifiers to predict whether a pair of words are translations of one another or not. The translations in our training data serve as positive training examples. The negative training examples are constructed by randomly pairing source language words in the training data with English words.[4] We use our development data to set the number of negative examples positive example. Using three negative examples for each positive example optimized performance on the development set. At test time, after scoring all source language words in the test set paired with all English words in our candidate set,[5] we rank the English candidates by their classification scores and evaluate accuracy in the top-k translations.

We use the fast, online learner implemented in the Vowpal Wabbit package (Agarwal *et al.*, 2014) to estimate the parameters of our log-linear classifiers.[6] VW uses a gradient descent-based algorithm for learning binary predictors, and we perform 100 learning passes over the training data. Although our current feature space is somewhat small,

[4] Among those that appear at least ten times in our monolingual data, consistent with our candidate set.

[5] All English words appearing at least three times in our monolingual data. In practice, we further limit the set to those that occur in the top-1000 ranked list according to at least one of our signals. Because words outside of these top-1000 lists are extremely unlikely to end up with a relatively high prediction score, doing so does not impact our performance but speeds up the prediction step.

[6] We use `http://hunch.net/~vw/` version 6.1.4, and run it with the following arguments that affect how updates are made in learning: –exact adaptive norm –power t 0.5

src	trg	A	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
politic	political	T	0.127	0.000	2.000	0.165	0.139	0.722	0.644	0.134	0.359	0.891	0.000	0.000	0.465	0.179	0.000	0.000	0.000	0.095
	offing	F	0.000	0.879	6.500	0.000	0.000	7.414	0.391	0.027	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.402
	first	F	0.000	0.000	6.000	0.000	0.130	2.490	0.274	0.239	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.133	0.000	0.081
	shipbuilding	F	0.161	0.000	9.500	0.000	0.000	3.358	0.638	0.072	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.155
curs	course	T	0.000	0.000	2.000	0.000	0.055	0.437	0.820	0.036	0.000	0.000	0.000	0.000	0.000	0.052	0.000	0.000	0.000	0.107
	refresher	F	0.092	0.000	6.500	0.000	0.000	7.132	0.380	0.031	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.369
	meeting	F	0.089	0.000	5.500	0.000	0.000	0.702	0.933	0.033	0.175	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.110
	pirc	F	0.000	0.000	3.000	0.000	0.000	7.374	0.358	0.038	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.402
valea	valley	T	0.000	0.925	2.000	0.000	0.000	0.036	0.693	0.184	0.000	0.000	0.000	0.919	0.000	0.000	0.000	0.000	0.000	0.103
	geography	F	0.000	0.000	7.000	0.000	0.012	0.074	0.509	0.377	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.102
	either	F	0.000	0.000	5.500	0.000	0.013	0.250	0.566	0.056	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.100
	birthday	F	0.000	0.908	6.500	0.000	0.000	1.785	0.994	0.049	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.126
olanda	netherlands	T	0.194	0.000	8.500	0.293	0.000	0.218	0.805	0.247	0.349	0.000	0.000	0.000	0.315	0.000	0.000	0.000	0.000	0.107
	vows	F	0.121	0.000	5.000	0.000	0.000	3.396	0.691	0.065	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.163
	orava	F	0.000	0.000	3.000	0.000	0.000	5.337	0.499	0.759	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.237
	kunduz	F	0.000	0.000	6.000	0.235	0.000	5.415	0.471	0.688	0.000	0.000	0.000	0.000	0.255	0.000	0.000	0.000	0.000	0.241
revista	magazine	T	0.000	0.000	7.500	0.208	0.000	0.028	0.726	0.405	0.000	0.000	0.000	0.000	0.338	0.050	0.178	0.040	0.000	0.105
	takwin	F	0.603	0.000	6.500	0.000	0.000	8.167	0.000	0.000	0.000	0.000	0.061	0.000	0.000	0.000	0.000	0.000	0.000	1.000
	archeological	F	0.065	0.000	10.000	0.000	0.000	2.832	0.771	0.373	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.149
	hollie	F	0.000	0.000	6.500	0.000	0.047	7.231	0.432	0.109	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.417
adus	brought	T	0.398	0.000	5.500	0.260	0.091	0.311	0.630	0.428	0.329	0.000	0.378	0.000	0.000	0.091	0.000	0.000	0.000	0.104
	centuryfrom	F	0.344	0.000	7.500	0.000	0.000	7.982	0.000	0.000	0.246	0.960	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
	associated	F	0.000	0.000	7.000	0.000	0.059	0.170	0.681	0.536	0.000	0.959	0.000	0.000	0.000	0.074	0.000	0.062	0.000	0.105
	abuse	F	0.000	0.000	2.000	0.000	0.000	1.591	0.875	0.407	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.129

Table 4.10: Examples of Romanian and English candidate word pairs scored with all 18 features. Feature numbers correspond to those enumerated in Section 4.2.3. The third column, 'A,' indicates if the words are true translations (T) or not (F).

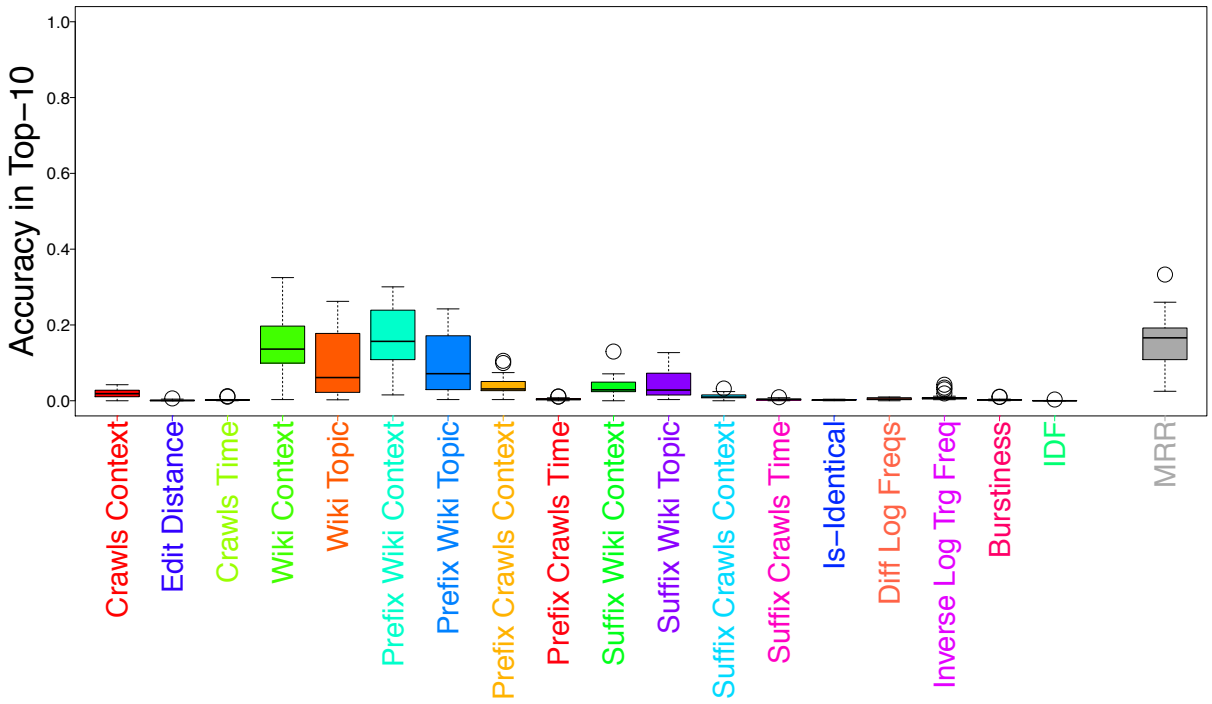


Figure 4.3: Performance using each of the 18 features separately to rerank translation candidates as well as our baseline method for combining them, which uses the simple mean reciprocal rank across all features. Box and whisker plots depict the distribution of performance across a set of 24 languages. The three lines in each box illustrate the first, second (median), and third quartiles. Outliers (defined as being more than 1.5 times the interquartile range away from either quartile) are shown with circles. The whiskers show non-outlier minimum and maximum values.

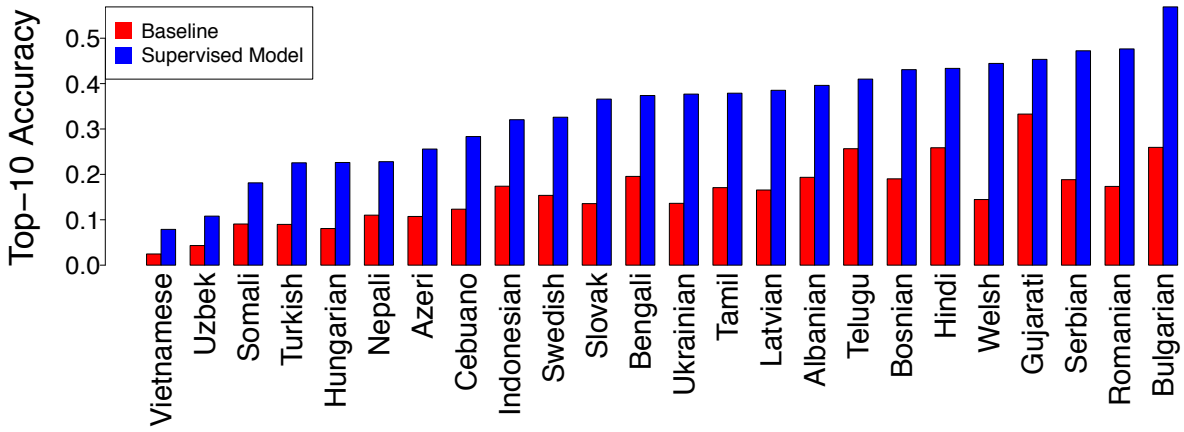


Figure 4.4: Top-10 bilingual lexicon induction accuracy of the baseline MRR approach to combining signals and our proposed supervised approach for each of 24 languages.

in future work we plan to scale up learning to take advantage of, for example, full context vectors, and VW will be well-suited to learn over very large feature spaces. We train classifiers separately for each source language, and the learned weights vary based on, for example, corpora size and the relatedness of the source language and English (i.e. the number of cognates). Although the scale of feature values varies somewhat (e.g. frequency difference can be greater than 1), making it difficult to interpret feature weights, we compared feature weights and found that the highest weighted feature for 19 languages is the Wikipedia topic similarity feature, and the highest for 5 languages is the Wikipedia context feature. These results are consistent with what we saw comparing the performance of individual features in Figure 4.3.

4.3 Experiments

For each source language, we use our trained models to induce translations for each source language word in our test sets, and we do evaluation against our gold standard bilingual dictionaries. We rank English translations by their translation classification score and measure percent accuracy in the top-k. This measure is somewhat conservative since the dictionaries aren't expected to be exhaustive, meaning that some target language translations for a given source language word won't appear in the dictionary and the system won't be given credit for ranking these target items high in its translation list. This is particularly true here because we have used the MTurk dictionaries, which are somewhat noisy. However, in these experiments, we only evaluate on words that do appear in our bilingual dictionary. It's possible that such words are easier to translate than, say, a given OOV word in some sentence which we wish to translate. The results presented in this section are on the held-out blind test sets described above.

4.3.1 Comparison with Unsupervised Baseline

Table 4.11 shows the top-10 bilingual lexicon induction accuracy results for each language using the baseline model as well as the supervised discriminative models.[7] Figure 4.4 shows the same top-10 accuracies, sorted by performance. Finally, Figure 4.5 shows summary box-and-whisker plots for both the MRR baseline and our proposed supervised model. It's clear that the supervised method outperforms the baseline by a large margin for all 24 languages. Results using the supervised models vary from 11% accuracy on Uzbek to 57% accuracy on Bulgarian. The average accuracy across languages using the MRR baseline is 15.8% and using a supervised approach is 34.2%, or greater than *twice* the average baseline accuracy.

[7]Performance is different from that reported in Irvine and Callison-Burch (2013b) because, here, we have done additional quality control and cleaning over our MTurk dictionaries, which we use for training and evaluation, as described in 3.2.2. Our feature set is also slightly different, including, for example, the burstiness feature, which was not a part of our original feature set presented in Irvine and Callison-Burch (2013b).

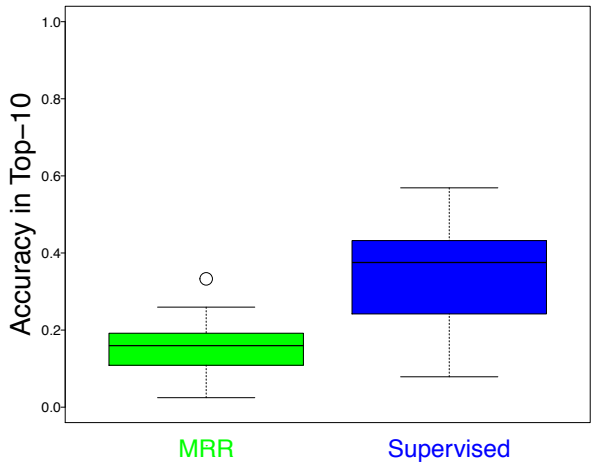


Figure 4.5: Summary box-and-whisker plots of the top-10 bilingual lexicon induction accuracies of the baseline and our proposed supervised approach over 24 languages

4.3.2 Analysis by Word Frequency

Figure 4.6 presents results on the same set of experiments, but bins source language words by their Wikipdia corpus frequency.[8] We binned the words in each evaluation test set by frequency, and each bin contains 100 source language words. That is, the most frequent 100 source language words were put in the first bin, and the least frequent were put into the last bin. The horizontal axis in each figure plots the average corpus frequency of the words in a given bin versus the percent of those source language words that have a correct translation in the top-k ranked list of translations.

The results in Figure 4.6 are presented starting with the language with the least amount of Wikipedia data (Somali) and ending with the language with the largest amount (Swedish), among those languages for which results are presented. Corpus frequencies for even the most frequent words in the first few source languages are very small. For example, the average frequency of the 100 most frequent Somali words is only 13.

Prior work on bilingual lexicon induction has focused on identifying translations for frequent words. In general, our monolingual signals are stronger for those words that appear frequently in monolingual corpora than for those words that appear less frequently and have sparse context and temporal counts. Therefore, we hypothesized that translation accuracy would be higher for frequent words than for less frequent words, resulting in accuracies that go up from left to right, or from lower frequency to higher frequency, in the figures. Figure 4.6 shows that this effect holds true, but it is not as strong as we expected.

To quantify the effects of frequency, we compute the Spearman rank-order correlation coefficient between the frequency rank of a given source word and the rank of its correct translation.[9] Across all languages, we find a slightly positive average correlation on average, 0.08, indicating that, as we expected, more frequency words tend to have higher ranked correct translations. This effect is significant to a p-value of 0.01 for 14 of the 24 languages,[10] however the correlation is not as large as we expected. In Section 4.3.3 we do a similar analysis based on burstiness.

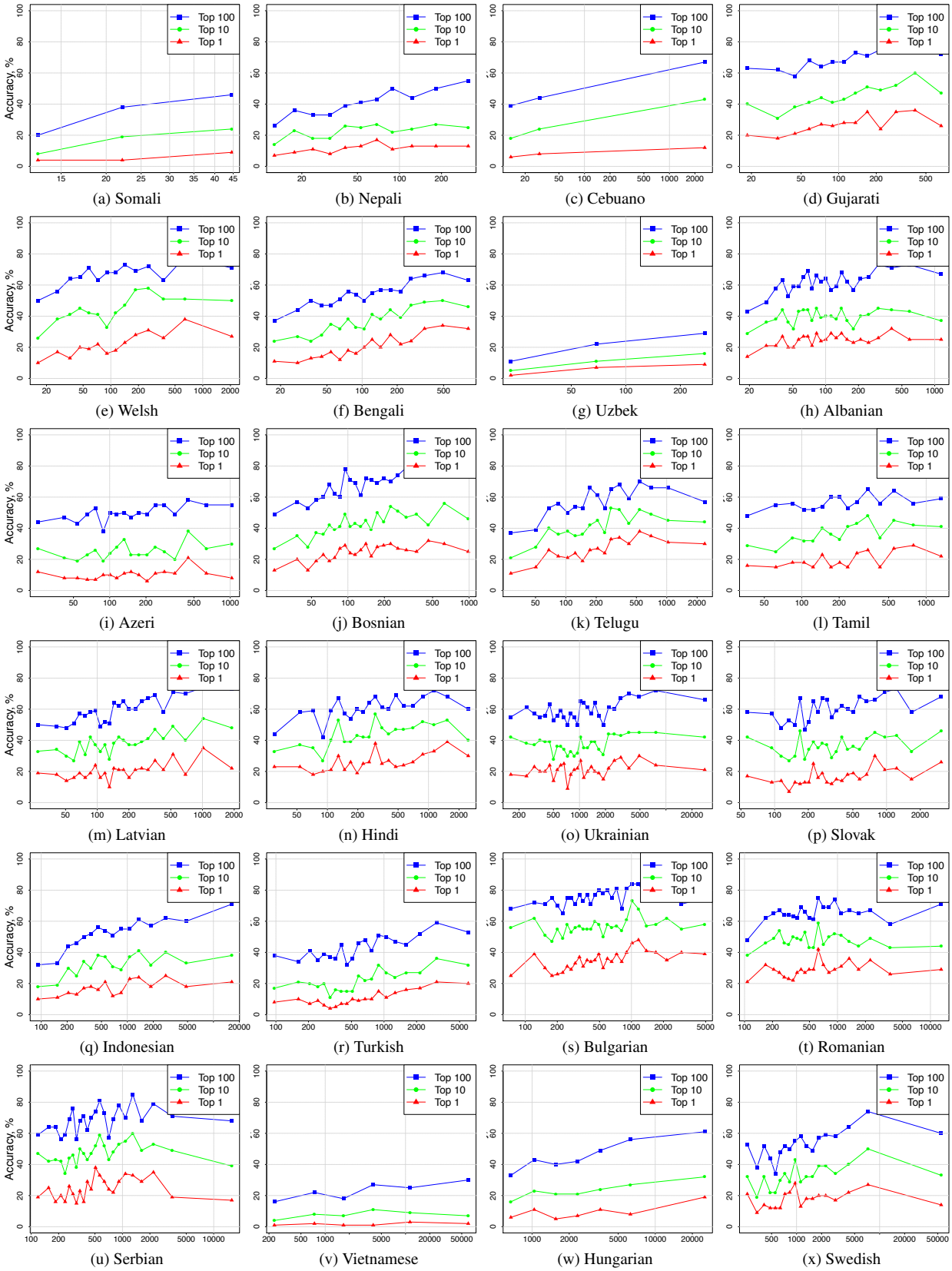
4.3.3 Analysis by Word Burstiness

Figure 4.7 presents results again on the same set of experiments but bins source language words by their Wikipedia corpus *burstiness*. We use the burstiness definition (B_w, not IDF_w) given in Section 4.2.1.6. As we did for the word frequency analysis, we bin the words in each evaluation set by burstiness, with each bin containing 100 source words. That is, the 100 most bursty source language words were put in the first bin, and the least bursty were put into the last

[8] Because the features estimated over the Wikipedia corpora are much stronger than those estimated over the web crawls, we find the analysis over Wikipedia frequency to be more meaningful than combining word frequencies across corpora.

[9] Although we have integer-valued frequency information, our comparison variable only contains ranks, so we convert frequency to an ordinal variable by ranking the words in each test set by their Wikipedia monolingual frequencies, from highest to lowest.

[10] Bosnian, Cebuano, Somali, Nepali, Gujarati, Bengali, Latvian, Indonesian, Welsh, Tamil, Turkish, Telugu, Hungarian, Swedish



(a) Somali (b) Nepali (c) Cebuano (d) Gujarati

(e) Welsh (f) Bengali (g) Uzbek (h) Albanian

(i) Azeri (j) Bosnian (k) Telugu (l) Tamil

(m) Latvian (n) Hindi (o) Ukrainian (p) Slovak

(q) Indonesian (r) Turkish (s) Bulgarian (t) Romanian

(u) Serbian (v) Vietnamese (w) Hungarian (x) Swedish

Figure 4.6: Bilingual lexicon induction as a function of source word **frequency** in Wikipedia monolingual data. Among the languages shown, we have the least monolingual data for Somali and the most for Swedish.

Language	MRR Baseline	Supervised Model	Absolute Improvement	% Relative Improvement
Vietnamese	2.5	7.9	5.4	216.0
Uzbek	4.3	10.8	6.5	151.2
Somali	9.1	18.1	9.0	98.9
Turkish	9.0	22.5	13.5	150.0
Hungarian	8.1	22.6	14.5	179.0
Nepali	11.0	22.8	11.8	107.3
Azeri	10.7	25.6	14.9	139.3
Cebuano	12.3	28.3	16.0	130.1
Indonesian	17.4	32.0	14.6	83.9
Swedish	15.4	32.6	17.2	111.7
Slovak	13.6	36.6	23.0	169.1
Bengali	19.6	37.4	17.8	90.8
Ukrainian	13.6	37.7	24.1	177.2
Tamil	17.1	37.9	20.8	121.6
Latvian	16.6	38.5	21.9	131.9
Albanian	19.4	39.6	20.2	104.1
Telugu	25.7	41.0	15.3	59.5
Bosnian	19.0	43.1	24.1	126.8
Hindi	25.9	43.4	17.5	67.6
Welsh	14.5	44.4	29.9	206.2
Gujarati	33.3	45.3	12.0	36.0
Serbian	18.8	47.2	28.4	151.1
Romanian	17.3	47.6	30.3	175.1
Bulgarian	26.0	56.9	30.9	118.8
Average	15.8	34.2	18.3	129.7

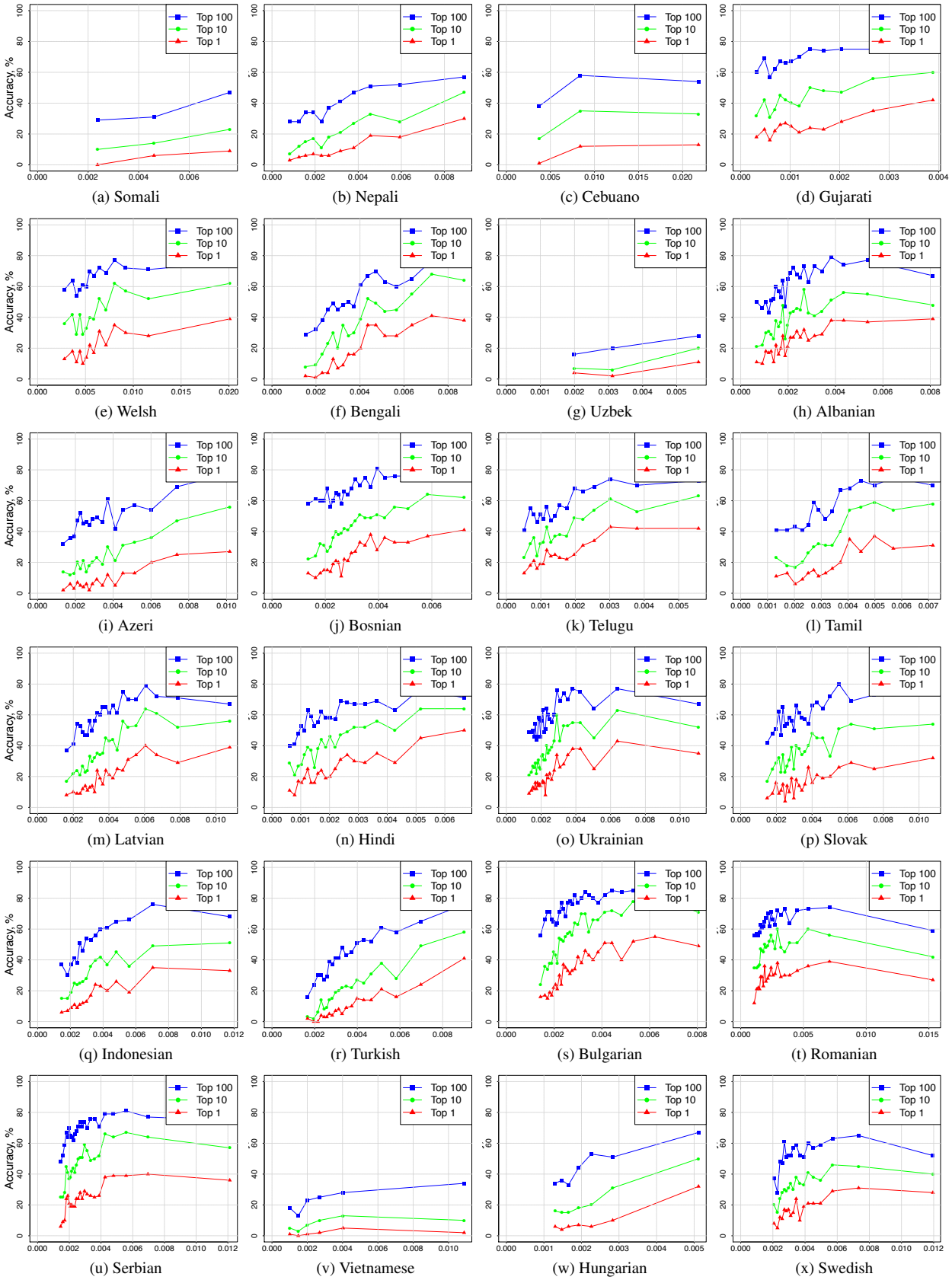
Table 4.11: Top-10 Accuracy on test set. Performance increases for all languages moving from the baseline (*MRR Baseline*) to discriminative training (*Supervised Model*). The average accuracy across languages using the MRR baseline is 15.8 and using our supervised approach is 34.2.

bin. The horizontal axis in each figure plots the average burstiness of the words in a given bin versus the percent of those source language words that have a correct translation in the top-k ranked list of translations.

We hypothesized that it may be easier to induce translations for bursty words than for non-bursty words because their temporal and topic signatures are very peaked. The results in Figure 4.7 confirm this. Again, without binning by burstiness, we compute the Spearman rank-order correlation coefficient between the rank of a given word's burstiness and the rank of its correct translation. Across all languages, we find a positive average correlation on average, 0.25, indicating that, as we expected, we tend to rank correct translations higher for more bursty words. This effect is significant to a p-value of 0.01 for *all* 24 languages. Comparing our results here with those in Section 4.3.2, we see that burstiness is a better indicator of ranking performance on a given word than frequency.

4.3.4 Analysis by Amount of Monolingual Data

Figure 4.8 plots the average top-10 and top-100 accuracies versus the total amount of monolingual data (web crawls and Wikipedia; amounts given separately in Table 11.3 in Appendix 11) for each of the 24 languages. In general, an increase in monolingual data seems to improve accuracy. The correlation is not perfect, however. For example, performance on Turkish and Vietnamese is relatively poor despite the relatively large amount of monolingual data available for each.



(a) Somali (b) Nepali (c) Cebuano (d) Gujarati

(e) Welsh (f) Bengali (g) Uzbek (h) Albanian

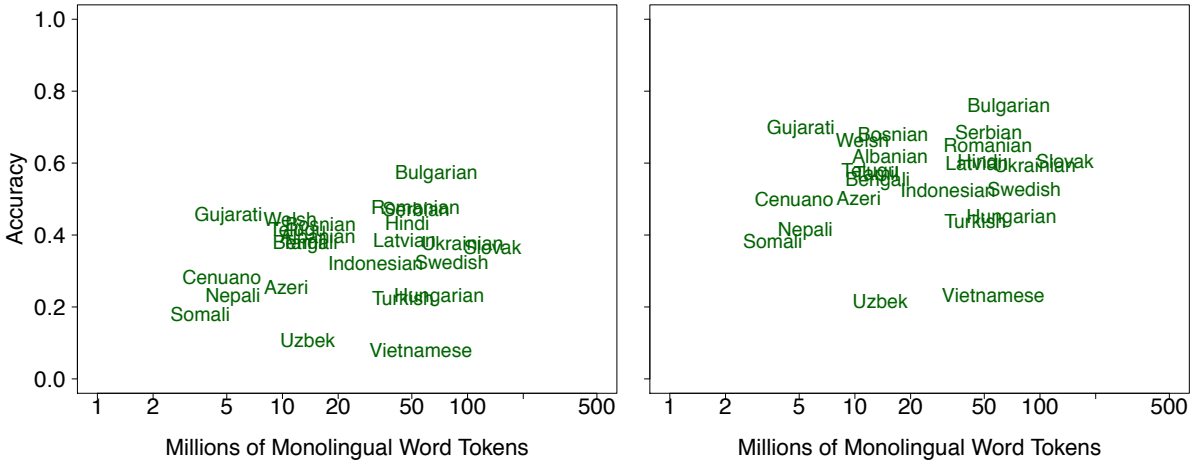
(i) Azeri (j) Bosnian (k) Telugu (l) Tamil

(m) Latvian (n) Hindi (o) Ukrainian (p) Slovak

(q) Indonesian (r) Turkish (s) Bulgarian (t) Romanian

(u) Serbian (v) Vietnamese (w) Hungarian (x) Swedish

Figure 4.7: Bilingual lexicon induction as a function of source word **burstiness** in Wikipedia monolingual data.



(a) Accuracy in Top-10

(b) Accuracy in Top-100

Figure 4.8: Total size of source language comparable corpora (Wikipedia and web crawls) in millions versus top-10 bilingual lexicon induction accuracy.

4.4 Learning Curve Analyses

4.4.1 Translated Word Pairs

Figure 4.9 shows learning curves over the number of positive training instances for each source language. In all cases, the number of randomly generated negative training instances is three times the number of positive. For all languages, performance is stable after about 300 correct translations are used for training. This shows that our supervised method for combining signals requires only a small training dictionary. In most cases, for a new language, a dictionary of this size could be mined from the Internet or created using crowdsourcing (Irvine and Klementiev, 2010).

4.4.2 Monolingual Data

Given the results presented in Section 4.3, one remaining question that we may want to answer is the following: How much monolingual data would we need to ensure high quality induced bilingual lexicons? Or, put slightly differently, do our experiments show any signs of bilingual lexicon induction performance leveling off after a certain amount of monolingual data is available? If so, any further performance gains would have to be made by improving our techniques instead of, for example, expanding our web crawls to additional websites. These are important considerations as we move to integrating induced translations into end-to-end SMT.

Figure 4.10 shows bilingual lexicon induction learning curves for four languages, Gujarati, Albanian, Azeri, and Tamil. Top 1, top 10, and top 100 accuracies are plotted on the y-axis for each language, and the x-axis shows the amount of monolingual data used to score and rank translation candidates. We generated the learning curves by sampling the web crawl and wikipedia monolingual corpora at the same rate. The total amount of monolingual data available for Gujarati is about 5 million words, and it is about 11 million for Azeri, 13 million for Tamil, and 15 million for Albanian.

Performance levels off after about half of the Azeri and Tamil data and one third of the Albanian data are used. This corresponds to about 5 million words. For Gujarati, performance increases rapidly up to the full amount of 5 million monolingual words. These results indicate that we need at least a few million words of comparable corpora to achieve good performance, and using more monolingual data does not harm performance.

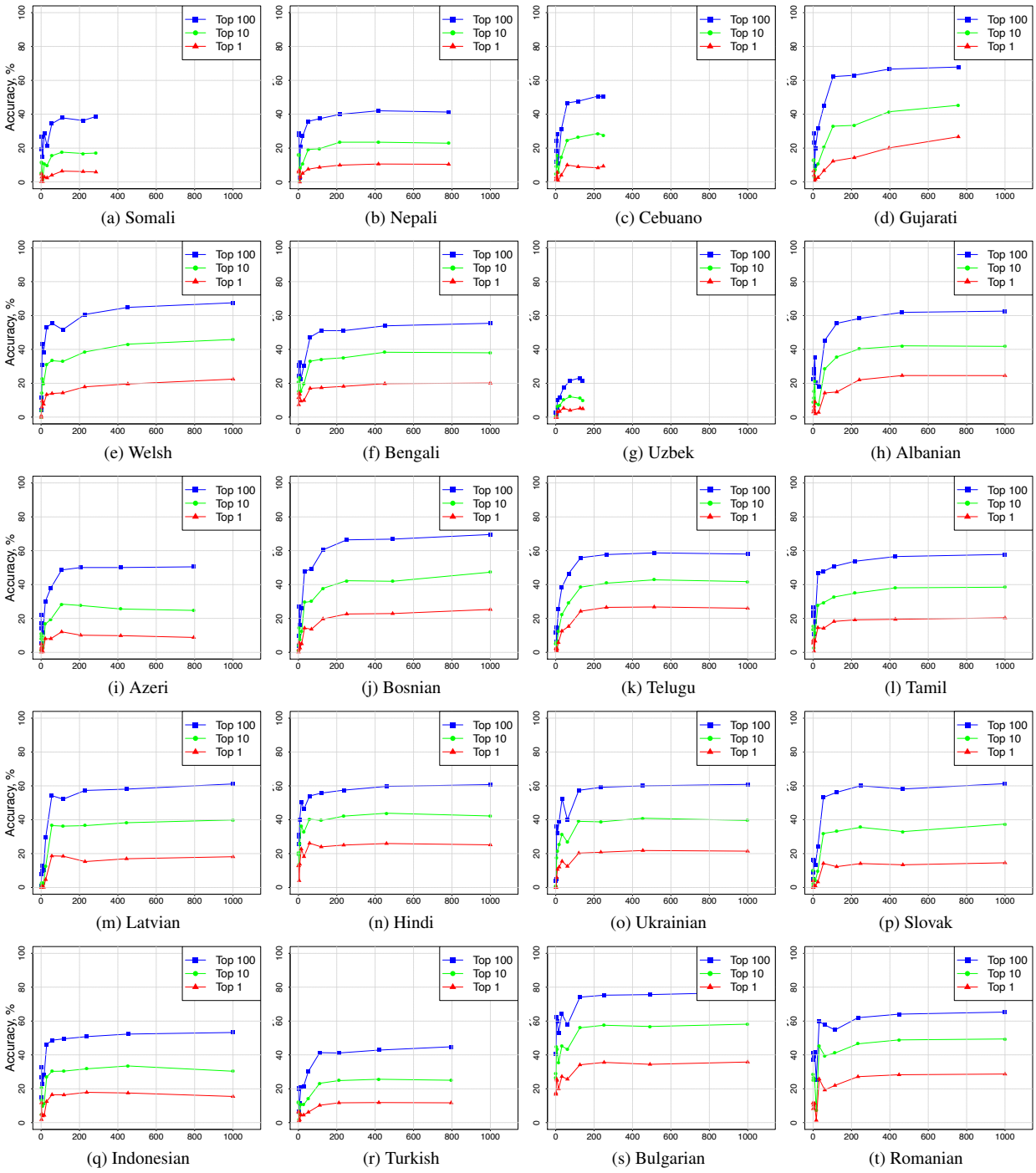
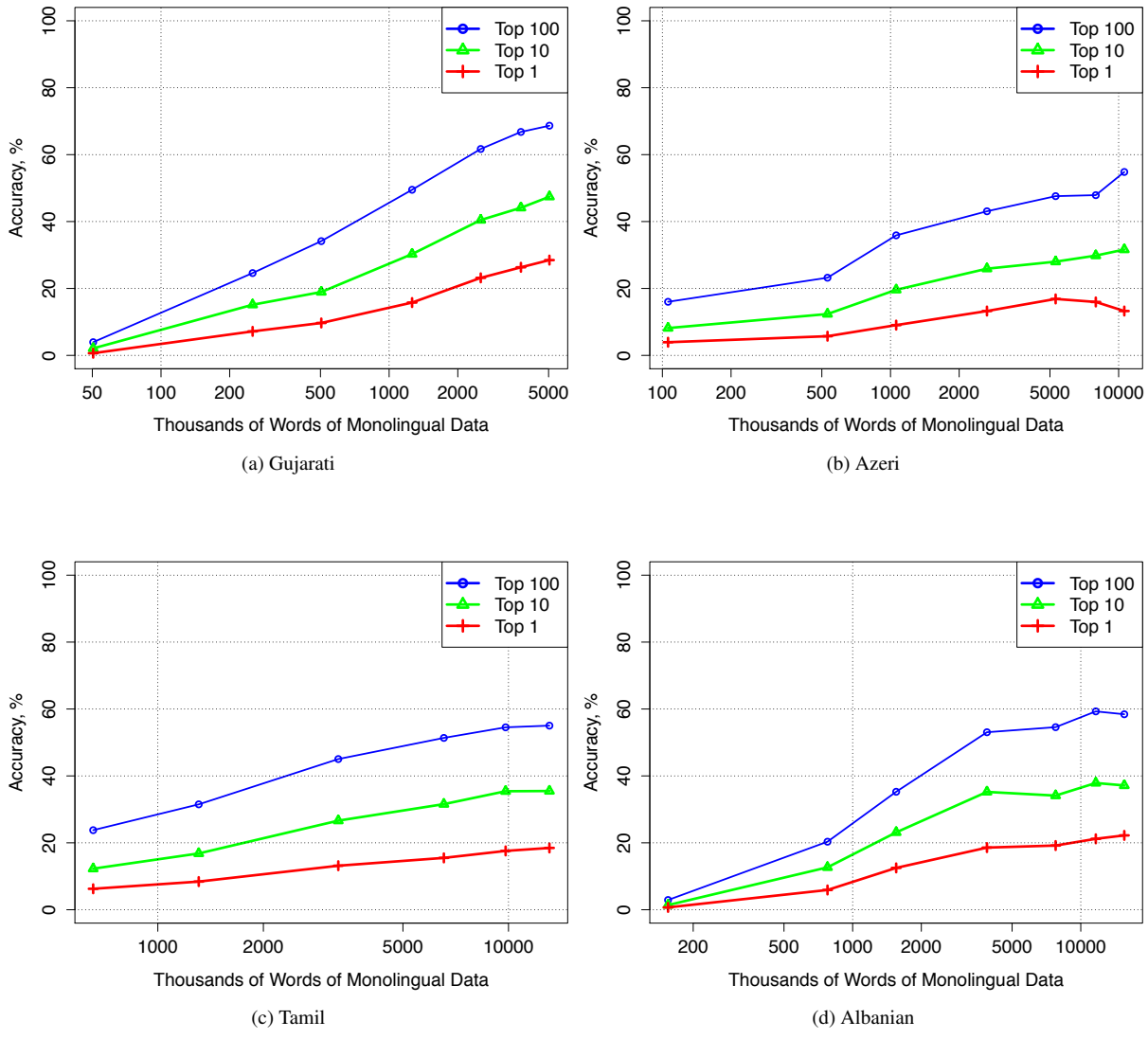


Figure 4.9: Learning curves over number of positive training instances, up to 1,000. For some languages, 1,000 positive training instances are not available. In all cases, the number of negative training instances is three times the number of positive. For all languages, performance is fairly stable after about 300 positive training instances.



(a) Gujarati

(b) Azeri

(c) Tamil

(d) Albanian

Figure 4.10: Bilingual lexicon induction learning curves over varying comparable corpora sizes for (a) Gujarati, (b) Albanian, (c) Azeri, and (d) Tamil. The x-axis is shown on a log scale.

4.5 Learning Models Across Languages

The results in Section 4.4.1 showed that we only need a few hundred pairs of translated words to learn a high-performing discriminative model for inducing word translations. Although we expect that this amount of data could be quickly gathered for any language pair of interest, this assumption may not always hold. It may be desirable then to use a model trained on data for another language pair. For example, if we were unable to obtain the few hundred Gujarati-English word translations needed to achieve high bilingual lexicon induction performance, we may use the classification model trained on Hindi data, which is a closely related language. Here, we present experiments using a model trained on data from one language pair to induce translations for another language pair. That is, for example, we test the effectiveness of the discriminative model *weights* that we learned using Hindi-English data to score and rerank hypothesis Gujarati-English translations, for example. Note that we only use the learned discriminative weights across languages (one for each feature); in all cases, we estimate the feature values themselves using same-language comparable corpora. In our example, we estimate feature values for pairs of Gujarati and English words using Gujarati-English comparable corpora, but then we use the weights learned over our Hindi training data to combine the features in order to predict Gujarati translations. In all experiments, we use the same set of 18 features described in Section 4.2.3.[11]

Interestingly, we find that the weight vectors obtained from other languages often result in higher test set performance than those obtained on that languages's development set. Table 4.12 shows the results for 20 language pairs. Results on the diagonal are identical to those presented in Table 4.11. Test set languages are listed in each row and columns indicate the trained model used to make predictions. For 6 of the 20 languages (Telugu, Tamil, Indonesian, Bulgarian, Romanian, Vietnamese), using the model trained on that language's training data performs better than using any of the 19 models trained using data in another language. These six languages are mostly relatively high resource languages, for which we have large comparable corpora. Interestingly, of the remaining 14 languages, for half (Nepali, Welsh, Bengali, Albanian, Azeri, Bosnian, Turkish) we see the highest performance on the test set when we make predictions using the model trained on Indonesian data. There do not seem to be any trends that correspond to language relatedness. In most cases, models trained on languages for which there is more comparable corpora available than there is for the test language tend to perform better. This is not an intuitive result because, at test time, features are computed over the test language's comparable corpora.

In general, Figure 4.12 shows that there is little variance in performance when we vary the model used to make predictions at test time. That is, the weights that we learn for our 18 features do not vary tremendously across the training datasets for different languages. This is an important finding as it suggests that, even if we do not have access to any example word translations for a given language pair, we may use a model trained on translations and comparable corpora from an alternate language pair to make high quality predictions.

4.6 Comparison with Prior Work

We compare our discriminative bilingual lexicon induction approach with the popular generative model presented in Haghighi *et al.* (2008). Haghighi *et al.* (2008) presents a canonical correlation analysis (CCA) based approach to inducing bilingual lexicons. The generative model presented in that work first generates a set of one-to-one matchings, M, between pairs of source and target words. Then, a feature vector is generated for each matched word type, s_i and t_j, from a 'language-independent concept,' $z_{i,j}$. Similar to our work, source and target words are represented by feature vectors characterizing their orthographies and contexts in monolingual corpora. However, unlike our work, the generative model proposed in Haghighi *et al.* (2008) allows neither source nor target word types to have multiple translations. Inference is done through boostrapped EM; the best CCA parameters, θ, are computed in the M-step, and the maximum weighted bipartite matching is found in the E-step using the Hungarian algorithm. In the first iteration, an initial lexicon is used to seed the E-step, and in additional EM iterations, an increasing number of high-confidence matchings are included until a complete bipartite matching is identified. The approach is referred to as matching canonical correlation analysis (MCCA).

In the original paper, Haghighi *et al.* (2008) presents results on several language pairs. However, evaluation is only done over nouns, which is a bursty word class, and lexicons are limited to high-frequency words. As we showed in Sections 4.3.2 and 4.3.3, frequent and bursty words tend to be the easiest to translate accurately.

Using code released by Haghighi *et al.* (2008), we directly compare our system with the MCCA approach. We

[11] Our contextual similarity feature is also dependent on having access to a bilingual dictionary for contextual vector projection. We ignore this here in order to keep results comparable to those presented in Section 4.3.

Train \ Test	so	ne	ceb	gu	cy	bn	uz	sq	az	bs	te	ta	lv	hi	sk	id	tr	bg	ro	vi	avg	st.dev.
so	18.1	20.0	13.3	18.1	19.5	18.7	18.1	19.5	16.3	21.6	19.5	17.3	18.9	17.6	17.9	21.1	**21.9**	17.3	18.9	14.1	18.4	2.1
ne	17.4	22.8	14.7	22.5	23.4	21.7	19.8	23.0	17.9	23.7	23.2	20.1	22.5	19.6	20.8	**25.6**	23.6	20.0	22.3	19.2	21.2	2.5
ceb	29.0	24.7	28.3	22.3	22.7	25.0	**29.7**	29.0	24.7	25.7	23.0	22.7	28.3	27.0	26.0	25.7	27.0	24.7	29.3	26.7	26.1	2.3
gu	27.7	43.0	27.8	45.3	43.0	43.9	38.5	40.5	39.7	43.8	**47.6**	43.8	43.5	43.6	41.3	43.1	41.5	43.3	39.2	45.3	41.3	5.0
cy	24.5	41.5	12.8	35.8	44.4	38.6	30.1	41.1	28.7	41.7	39.3	34.3	39.6	33.5	35.2	**45.2**	42.0	35.1	37.7	27.1	35.4	7.6
bn	29.9	36.8	27.3	35.2	34.5	37.4	33.6	36.9	36.2	36.0	37.0	37.2	37.1	36.4	36.2	**38.8**	35.6	37.1	34.1	35.9	35.5	2.6
uz	9.9	9.9	7.7	6.5	10.2	10.8	10.8	11.1	10.8	11.1	9.0	11.4	**12.3**	11.4	12.0	9.9	10.8	11.7	10.2	10.5	10.4	1.4
sq	29.3	38.9	18.5	36.3	39.6	34.1	33.7	39.6	31.5	38.0	37.9	30.1	38.3	32.6	34.4	**41.5**	39.0	33.9	40.4	31.7	35.0	5.2
az	18.3	25.8	15.5	23.6	26.3	27.3	22.9	26.2	25.6	25.3	27.3	27.3	26.9	26.0	25.4	**28.0**	25.8	27.4	23.2	24.1	24.9	3.0
bs	31.0	43.4	21.3	39.8	45.1	41.4	38.7	43.6	36.7	43.1	41.7	39.1	42.9	38.8	40.5	**47.0**	43.1	40.5	42.1	38.6	39.9	5.4
te	23.3	35.9	24.5	38.8	38.8	38.8	34.5	36.0	36.9	37.8	**41.0**	39.3	38.3	37.8	37.6	37.6	35.8	38.5	36.2	39.5	36.3	4.4
ta	25.3	34.2	25.6	33.8	33.5	37.6	29.3	33.3	35.5	33.9	37.4	**37.9**	35.4	36.0	35.6	36.4	33.8	37.1	30.4	35.7	33.9	3.5
lv	25.4	35.9	17.2	32.2	**39.0**	36.5	35.1	37.5	33.6	37.8	36.1	34.2	38.5	34.2	35.2	38.5	36.4	36.8	36.8	33.6	34.5	4.9
hi	31.8	41.9	28.9	42.5	39.7	43.6	37.2	40.4	40.9	41.0	44.1	44.1	42.0	43.4	41.8	43.0	40.8	43.4	38.3	**44.8**	40.7	4.0
sk	24.9	34.3	14.2	29.6	**39.0**	35.4	31.6	37.6	31.4	36.9	35.6	33.2	37.4	34.1	36.6	36.5	36.9	36.8	36.5	32.8	33.6	5.5
id	16.4	26.6	4.2	28.1	26.6	22.8	16.2	24.0	18.3	23.3	27.0	19.4	23.1	21.4	19.7	**32.0**	25.6	21.4	22.5	25.6	22.2	5.7
tr	12.6	21.9	4.5	20.7	24.4	20.4	13.9	21.9	15.2	20.4	22.2	17.9	20.1	18.0	17.9	**26.3**	22.5	19.0	20.7	19.0	19.0	4.6
bg	40.1	53.8	30.4	51.4	54.4	56.5	48.4	54.3	52.1	54.4	54.2	55.3	55.9	54.5	55.2	55.5	53.9	**56.9**	52.1	55.6	52.2	6.2
ro	35.7	40.4	12.6	36.7	42.4	38.0	40.3	44.9	39.6	40.1	40.9	31.4	43.0	38.0	40.7	43.4	42.0	39.7	**47.6**	40.3	38.9	6.9
vi	3.0	4.8	1.8	4.8	5.6	6.2	4.4	5.4	6.6	5.1	5.1	6.4	5.7	6.6	5.6	5.9	5.3	7.1	5.9	**7.9**	5.5	1.3

Table 4.12: Using bilingual lexicon induction models trained on data from one language pair to score and rerank translations for another pair. The top-10 translation accuracy is shown for all pairs of 20 languages. Results on the diagonal are identical to those presented in Table 4.11. The final two columns show the average and standard deviation across all 20 training languages for each test language. The most accurate result for each test set is bolded.

Model	Accuracy (%)
MCCA	15.1
Discriminative Model w/ Context and Orth. Features Only	24.3
Discriminative Model w/ All Features	42.3

Table 4.13: Comparison of bilingual lexicon induction accuracies using (1) matching canonical correlation analysis (MCCA), (2) our supervised discriminative model using only contextual and orthographic features, and (3) our supervised discriminative model using our complete feature set. Accuracy is measured as the percent of test set translations that are correctly matched by each model's full bipartite matching.

present experiments on the Spanish-English language pair, taking monolingual corpora from our Wikipedia collection and bilingual lexicons from our MTurk dictionary. We randomly sample about $6,000$ Wikipedia page pairs, which contain about 5 million words of text in both languages. This amount of monolingual data is comparable to what was used in the experiments presented in Haghighi *et al.* (2008). We identify a bilingual dictionary of $1,100$ word translation pairs in the MTurk dictionary for which both the source and target lexicons are unique and all words appear in monolingual corpora greater than ten times. We randomly select 100 word pairs to serve as a seed lexicon in the MCCA approach and as training data in our discriminative approach, and we use the remaining $1,000$ word pairs as an evaluation set.

We use the standard learning parameters in the MCCA code released by Haghighi *et al.* (2008), which include ten iterations of bootstrapped EM and a context window of size four. The MCCA model uses both orthographic features and contextual features estimated over the Wikipedia monolingual corpora. We use MCCA to compute a full bipartite matching and measure accuracy over the complete test set of $1,000$ translation pairs.

We use the seed lexicon of 100 word pairs to train our supervised discriminative model. As before, we randomly select three times as many negative examples for training. We then use the learned model to score all words in the source test lexicon paired with all words in the target test lexicon. In order to make our results comparable, we follow Haghighi *et al.* (2008) and use the Hungarian algorithm (Kuhn, 1955) to find the best set of one-to-one bipartite matchings across the source and target lexicons, maximizing the total score across all matchings. We measure the performance of our discriminative model using, like Haghighi *et al.* (2008), only orthographic and contextual features. Then, we also measure performance when we add our topic, frequency, and burstiness similarity features to the model.

Table 4.13 shows the performance of each bilingual lexicon induction model. The MCCA approach correctly matches 15% of the $1,000$ test set pairs. Our discriminative approach using only orthographic and contextual similarity features correctly matches 24%. When we add our full feature set, our model achieves 42% accuracy. These results demonstrate that our discriminative model needs no more training data than is needed to seed a generative model like the one presented in Haghighi *et al.* (2008). This is consistent with our results in Section 4.4.1, where we showed that our models can achieve high accuracies on the bilingual lexicon induction task using only small amounts of supervision.

In addition to our discriminative model outperforming the MCCA generative model on the matching task, it has the added advantage of not being restricted to predicting 1:1 word translations. This is critical as, even for closely related language pairs, many words do not have a one-to-one correspondence across languages. One example from the domain adaptation setting is the French word *enceinte*. In medical contexts, it translates as *pregnant* in English, but in government contexts it translates as *place*, *house*, or *chamber* and in scientific contexts is translates most frequently as *enclosures*. We would not want to restrict models of bilingual lexicon induction to choosing only one sense, or one translation, for French *enceinte*. That is, the polysemy of words varies across languages and it is important to be able to account for this in any model of bilingual lexicon induction.

4.7 Conclusions

On average, we observe gains of *greater than 100%* over an unsupervised rank-combination baseline by using a seed bilingual dictionary and a diverse set of monolingual signals to train a supervised classifier. Using supervision for bilingual lexicon induction makes sense. In some cases a dictionary is already assumed for computing contextual similarity, and, in the remaining cases, one could be compiled, for example through crowdsourcing (Callison-Burch and Dredze, 2010; Irvine and Klementiev, 2010; Pavlick *et al.*, 2014). Our supervised framework has the additional advantage that any new monolingually-derived similarity metrics can easily be added as new features.

Our experiments showed that we only need a few hundred example translations to learn a high quality model of bilingual lexicon induction. Additionally, we showed that even in the case that this small amount of supervision is not available, we can effectively use a model trained on a different language pair to induce translations. We have also observed that translating less frequent and less bursty words is harder than translating more frequent, more bursty words. We found that using more monolingual data does not hurt performance, but high quality lexicons can be induced with just a few million words of comparable corpora. These findings inform our machine translation experiments, which follow in Chapters 6 and 8.

In Chapter 5, we expand the ideas presented in this chapter to score pairs of *phrases*, in addition to pairs of unigrams. This allows us to score all phrase pairs found in a phrase-based statistical machine translation phrase table and, in theory, induce *phrasal* translations in addition to unigram translations. Doing so introduces lots of new challenges because there are many more phrases than word types and because phrase counts are sparser than word counts. These challenges are explored in detail in Chapter 7.

Chapter 5

Monolingual Phrase Table Scoring

In this chapter, we begin with the idealization that a set of hypothesis phrase pairs is given, and we use comparable corpora to score this existing phrase table. We define features over phrase pairs that are based on comparable corpora and are similar to those proposed for unigrams in Chapter 4. Then, we present SMT experiments where we replace the standard bilingually estimated feature set with new features estimated over comparable corpora. We examine the degradation in translation performance when bilingually estimated translation probabilities are removed, and show that more than 50% of the loss can be recovered with our proposed feature set. We further show that our monolingual features add 1.5 BLEU points when combined with standard bilingually estimated phrase table features. Much of the work in this chapter was published in Klementiev *et al.* (2012). In Klementiev *et al.* (2012) we also propose a novel algorithm for estimating a lexicalized reordering model from comparable corpora. However, here and in the rest of this thesis, we only focus on lexical choice, not reordering.

5.1 Phrase Table Scoring

Here we extend the feature set defined for words in Chapter 4 to phrases. We use the following signals of translation equivalence for phrase pairs:

1. Web Crawls Contextual Similarity
2. Web Crawls Temporal Similarity
3. Wikipedia Contextual Similarity
4. Wikipedia Topic Similarity
5. Orthographic Similarity

For each phrase pair in our set, we measure *phrasal* contextual, temporal, and topic similarity and *lexical* contextual, temporal, topic, and orthographic similarity.

5.1.1 Phrasal Features

We estimate phrasal features exactly as we did for word pairs. That is, we gather contextual, temporal, and topic vectors for each source and target phrase in our set of phrase pairs using the corresponding side of our comparable corpora. Phrasal temporal and topic vector signatures are M-dimensional, where M is the number of days or topics in our corpus, and each element contains counts of the number of times the given phrase appeared in the data associated with each date or topic. As we did for words, we normalize temporal and topic signatures by dividing each count by the total count, and, again, we use cosine similarity to compare pairs of normalized vectors:

$$sim_{temp}(F, E) = \frac{F \cdot E}{||F||||E||},$$

where F and E are source and target language phrase temporal or topic signatures, respectively.

We also estimate phrasal contextual similarity. Contextual vectors are collected as before: we identify each appearance of source phrase f and target phrase e in our comparable corpora and collect counts of the words that appear in the context of each phrase. That is, although we collect contextual vectors for phrases, each component of the vectors corresponds to how often a particular *unigram* appears in its context. We do not collect counts of phrases that appear in the context of f and e because their counts are generally very sparse. As with words, we compute the value of the k-th component of f's contextual vector, f_k as follows:

$$f_k = n_{f,k} * (log(n/n_k) + 1)$$

where $n_{f,k}$ and n_k are the number of times s_k appears in the context of f and in the entire corpus, and n is the maximum number of occurrences of any word in the data. We project each component of the source language contextual vectors into the English vector space and then use cosine similarity to compare contextual vectors.

Because the word order internal to a phrase may vary across its translations, phrasal orthographic similarity is unlikely to be informative. For example, English *artificial intelligence* translates into Spanish as *inteligencia artificial*. Although the phrase pair contains two pairs of cognates with strong orthographic similarity, because their word order is opposite, measuring orthographic similarity on the phrase pair directly would not be effective without additional modeling of movement.

5.1.2 Lexical Features

We compute not only phrasal features but also *lexical* similarity features for each phrase pair. Our lexical equivalents of the phrase-level similarity scores are based on the similarity of individual words within the phrases. To compute these lexical similarity features, we average similarity scores over all pairs of words in the two phrases. For example, for the phrase pair *the artificial intelligence* and *la inteligencia artificial*, we compute the average similarity between all nine word pairs. By averaging over all word pairs, we avoid propagating word alignment errors while also effectively penalizing unaligned words. In many cases, we observed that, because individual words are more frequent than multiword phrases, the accuracy of the lexical features is higher than their phrasal equivalents. We compute lexical features for all of the signals enumerated above: contextual, temporal, topic, and orthographic.

5.2 Experiments with An Existing Phrase Table

We use the Spanish-English (high resource) and Urdu-English (low resource) language pairs to test our method for estimating translation phrase table features from comparable corpora. This allows us to compare our method against the normal bilingual training procedure. We expect bilingual training to result in higher translation quality because it is a more direct method for learning translation probabilities. After removing bilingually estimated translation features, our goal is to recover as much of the loss as possible using features estimated over comparable corpora.

For both language pairs we begin with phrase-based SMT models trained using the Moses framework (Koehn *et al.*, 2007). With the exception of maximum phrase length, which is set to 3 in our Spanish experiments, we use default values for all of the parameters. We learn lexicalized reordering models from the parallel training corpora and use them in all experiments.[1]

For Spanish, we use the full Europarl v5 parallel training corpus (Koehn, 2005). Our Spanish experiments use a trigram language model trained on the English side of the Europarl corpus using SRILM with Kneser-Ney smoothing. We tune feature weights (rerunning tuning for all experiments) using minimum error rate training (MERT) and a development bitext of 2,553 sentence pairs. We use the development and test data distributed in WMT shared task (Callison-Burch *et al.*, 2010).[2] The Spanish test set is translated newswire articles consisting of 2,525 single-reference sentence pairs.

We use the Urdu training/development/test corpora released by Post *et al.* (2012) (more details on this dataset are given in Chapter 6). Following that work, in our Urdu experiments we use a 5-gram language model trained on the English side of the training corpus. For each Urdu experiment, we rerun tuning using batch MIRA (Cherry and Foster, 2012).

In our Spanish-English experiments, we compare the effect of estimating the parameters of our model from two sets of comparable corpora, detailed in Table 5.1:

[1] In Klementiev *et al.* (2012), we present a method for also estimating a lexicalized reordering model from comparable corpora.

[2] Specifcally, *news-test2008* plus *news-syscomb2009* for dev and *newstest2009* for test.

	Europarl	Gigaword	Wikipedia
date range	4/96-10/09	5/94-12/08	n/a
shared dates	829	5,249	n/a
ES articles	n/a	3,727,954	59,463
EN articles	n/a	4,862,876	59,463
ES lines	1,307,339	22,862,835	2,598,269
EN lines	1,307,339	67,341,030	3,630,041
ES words	28,248,930	774,813,847	39,738,084
EN words	27,335,006	1,827,065,374	61,656,646

Table 5.1: Monolingual training data statistics for Spanish-English phrase table feature replacement experiments.

	Spanish-English phrase table	Urdu-English phrase table
Phrase pairs	3,093,228	1,529,084
Foreign phrases	89,386	478,943
English phrases	926,138	579,944
Foreign Unigrams	13,216	32,198
Average # translations	98.7	7.7
Foreign Bigrams	41,426	63,373
Average # translations	31.9	4.3
Foreign Trigrams	34,744	89,867
Average # translations	13.5	3.1

Table 5.2: Statistics about the Spanish-English and Urdu-English phrase tables used in feature replacement experiments. The numbers of unique unigrams, bigrams, and trigrams are given for each language along with the average number of translations for each.

- First, we treat the two sides of the Europarl parallel corpus as a comparable corpus, where the contextual and temporal distributions of phrases are *very* similar across languages.[3]
- Next, we estimate the features from truly comparable, non-parallel corpora. To estimate the *contextual* and *temporal* similarity features, we use the Spanish and English Gigaword corpora.[4] These corpora are substantially larger than the Europarl corpus, providing 27x as much Spanish and 67x as much English for contextual similarity, and 6x as many paired dates for temporal similarity. *Topical* similarity is estimated using Spanish and English Wikipedia articles that are paired with inter-language links.

To project context vectors from Spanish to English, we use our electronic Spanish-English bilingual dictionary (see Table 11.2 in Appendix 11). The context vectors for words and phrases incorporate co-occurrence counts using a two-word window on either side.

In our Urdu experiments, we only measure the effect of using truly monolingual datasets to estimate feature values. In particular, we use the news crawl (286 million words of Urdu) and Wikipedia (3.2 million words of Urdu) datasets that we described in Section 3.3.

Across all of our experiments, we keep our sets of phrase pairs consistent. The Spanish phrase table contains over 3 million phrase pairs extracted from the word-aligned parallel Europarl corpus, and the Urdu table contains about 1.5 million phrase pairs. We maintain the phrase pairs but drop their associated translation scores and then estimate replacement similarity scores over comparable corpora. The set of possible translations is constrained for each source phrase and therefore is likely to contain good translations. However, the average number of possible translations is high, especially for Spanish where they range from nearly 100 translations for each unigram to 14 for each trigram. For Urdu, because the training data is smaller, the average number of translations is smaller, ranging from 8 for each unigram to 3 for each trigram. Table 5.1 gives some data statistics about the set of phrase pairs for both source languages. The phrase tables contain a lot of noise, which results in poor end-to-end translation quality without good estimates of the translation quality of each phrase pair, as Section 5.2.1 shows.

5.2.1 Results

5.2.1.1 Ablation Experiments

Figure 5.1 shows the results of our Spanish experiments and Figure 5.2 shows the results of our Urdu experiments. Our standard SMT model for Spanish achieves a BLEU score of 21.87, and our standard model for Urdu achieves a BLEU score of 20.39. When we remove phrase table features from each, the Spanish and Urdu BLEU scores drop 9 points to to 12.86 and 8 points to 12.32, respectively.

Spanish Experiments 3-6 show how much our proposed features can help the model recover when they are estimated over the Europarl training corpus, treating the two sides as a comparable corpus. Of the temporal, orthographic, and contextual features, the temporal feature performs the best. Together (Experiment 6, 'All CC'), they recover more than each individually, yielding a total gain of 4 BLEU points.

Spanish Experiments 8-12 estimate each of the features from non-parallel comparable corpora. Remarkably, estimating our new features over the non-parallel corpora (Experiment 12) performs *better* than estimating features over the parallel corpus itself (Experiment 6). Using all of the features estimated over the Gigaword and Wikipedia corpora yields a total gain of over 5 BLEU points, or about 56% of the BLEU point loss that occurred when we dropped all of the original translation features. Much of this gain is due to the topic similarity feature, which we estimate using Wikipedia but for which we have no equivalent Europarl feature.

For Urdu, Experiments 3-7 show how using features estimated over monolingual data can help recover some of the loss incurred when we remove all features (Experiment 2). The context and topic features alone help more than either the temporal or orthographic feature. Using all four new feature types, we regain about 3 BLEU points, or about 36% of the BLEU point loss.

5.2.1.2 Combining Bilingually and Monolingually Estimated Features

Finally, we *supplement* the standard bilingually estimated model parameters with our monolingual features (Spanish Experiment 13, Urdu Experiment 8). For Spanish, we see a **1.5 BLEU** point increase over the standard model. Again,

[3] Haghighi *et al.* (2008) also used this method to show how well translations could be learned from monolingual corpora under ideal conditions, where the contextual and temporal distribution of words in the two monolingual corpora are practically identical.

[4] We use the Agence France-Presse (afp), Associated Press Worldstream (apw), and Xinhua News Agency (xin) sections of the corpus.

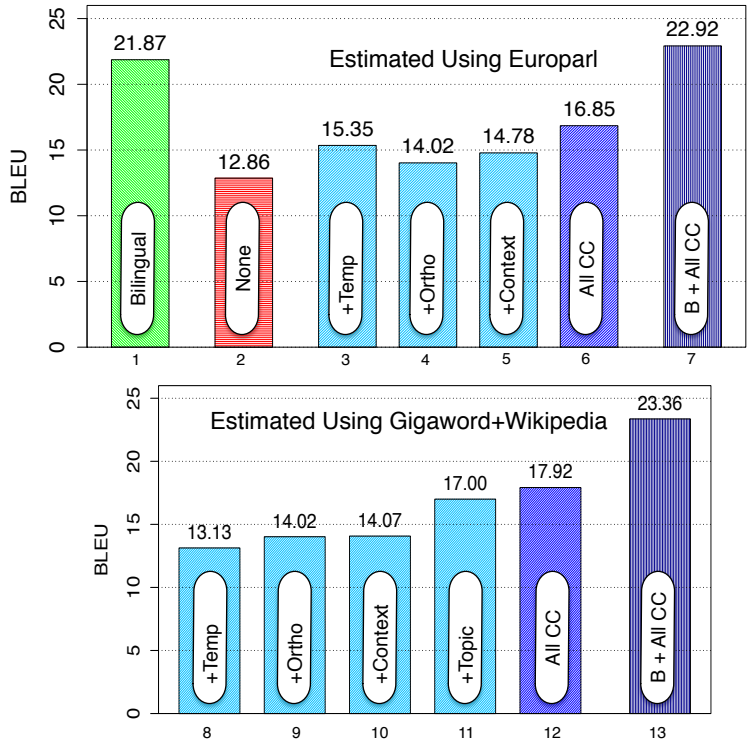


Figure 5.1: Spanish-English results replacing bilingually estimated phrase table features with monolingually estimated features. Much of the loss in BLEU score when bilingually estimated features are removed from a Spanish-English translation system (Experiment 2) can be recovered when they are replaced with monolingual equivalents estimated from Europarl data (Experiments 3-6; 'CC' refers to features estimated over comparable corpora). The second chart shows the performance of monolingual features derived from non-parallel comparable corpora. Over 56% of the BLEU score loss can be recovered (Experiment 12). When we use both bilingually and monolingually estimated features (Experiments 7 and 13), BLEU scores improve by over one point.

this result is higher than if we supplement the original model with features estimated over the parallel training corpus (Experiment 7). This indicates that our monolingually estimated scores as well as the Gigaword and Wikipedia comparable corproa are able to capture some novel information not contained in the standard feature set. For Urdu, we see a 0.6 BLEU point gain when we use monolingually estimated features in addition to the standard bilingually estimated ones.

5.3 Phrase Table Scoring Conclusions

This chapter showed that, given a high-quality but noisy phrase table, our new set of translation features, which are based on comparable corpora and are closely related to our bilingual lexicon induction feature set, do a good job of distinguishing good phrase pairs from bad phrase pairs. By comparing the full monolingual experiment ('All CC' in Figure 5.1) with the experiment that used the bare phrase table to translate, we can see that the monolingual scores greatly improve that baseline model's accuracy. Using the new features based on comparable corpora, we were able regain over 56% of the BLEU point loss that occurred when we dropped the bilingually estimated phrase table features.

We also observed promising results *supplementing* a baseline SMT model with features estimated over comparable corpora. In Chapter 6 we expand upon that idea and, beginning with a small bitext, use comparable corpora to supplement a baseline SMT model. There we also drop the idealization that we begin with a high quality phrase table. Later, in Chapter 8 we use the feature set described here to supplement SMT models in a domain adaptation setting.

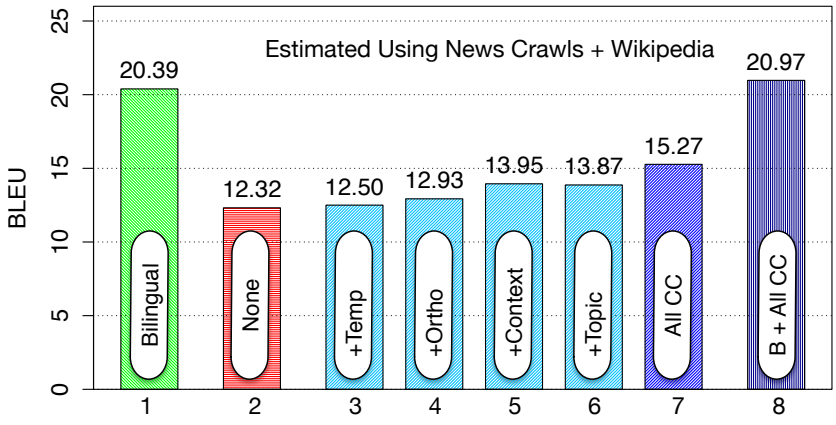


Figure 5.2: Urdu-English results replacing bilingually estimated phrase table features with monolingually estimated features. For Urdu, over 36% of the loss in BLEU score when bilingually estimated features are removed can be recovered when they are replaced with monolingual equivalents estimated from comparable corpora (Experiment 7). When we use both bilingually and monolingually estimated features (Experiments 8), the BLEU score improves by over half a point.

Chapter 6

End-to-End SMT with Zero or Small Parallel Texts

In this chapter, we build upon the bilingual lexicon induction technique developed in Chapter 4 as well as our approach to scoring phrase pairs using comparable corpora, which we presented in Chapter 5. Here, we consider the settings in which we have access to (1) bilingual dictionaries but no parallel sentences for training, and (2) only a small amount of parallel training data. In the first case, we wish to augment a baseline dictionary gloss with additional translations and features estimated using source and target language comparable corpora. Similarly, in the second case, we wish to augment a baseline statistical model learned over small amounts of parallel training data with additional translations and features estimated over comparable corpora. Assuming access to a bilingual dictionary or small amount of parallel text is realistic, especially considering the recent success of crowdsourcing translations (Ambati, 2011; Pavlick *et al.*, 2014; Post *et al.*, 2012; Zaidan and Callison-Burch, 2011).

We frame the shortcomings of SMT models trained on limited amounts of parallel text[1] in terms of accuracy and coverage. Coverage refers to the word and phrase translations that a model has any knowledge of at all, and it is low when the training text is small, which results in a high out-of-vocabulary (OOV) rate. Even for source language words and phrases that we do observe in small training sets, many of the possible target language translations are typically not observed. Our definition of coverage encompasses both SEEN and SENSE errors, which we introduced in Chapter 3. It can also loosely be thought of as the *recall* of a given SMT model, or, of all of the possible translations that exist between a pair of languages, the number which are included in the model.

Accuracy refers to the correctness, or the *precision*, of the translation pairs and their corresponding probability features that make up the translation model. Our SMT models contain both translation and reordering probabilities. In Chapter 3, we introduced SCORE errors, and we used a TETRA analysis to measure the impact of both translation SCORE and reordering SCORE errors in low resource settings. We found that because the quality of unsupervised automatic word alignments correlates with the amount of available parallel text and alignment errors result in errors in extracted translation pairs, accuracy tends to be low in low resource settings. Additionally, estimating translation and reordering probabilities over sparse training sets results in inaccurate feature scores.

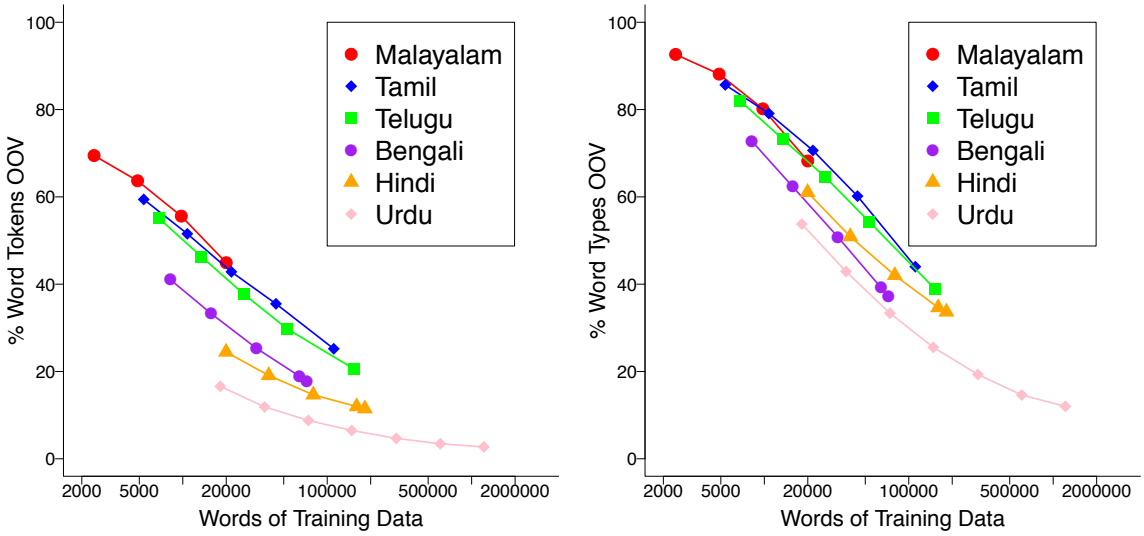
In this Chapter, we focus on SEEN errors, which impact the coverage of an SMT model, and translation SCORE errors, which impact the accuracy of an SMT model. By inducing translations for low frequency words, we also improve SENSE errors. We do not tackle reordering SCORE errors.

6.1 Improving Coverage

Figure 6.1 shows the percent of word tokens and word types in a development set that are OOV with respect to varying amounts of training data for several Indian languages (datasets described in Section 6.4.1[2]). In order to improve the coverage of our low resource translation models, we use the supervised bilingual lexicon induction technique that we presented in Chapter 4 to learn translations for words which appear in our test sets but not our training data (OOVs).

[1] We consider low resource settings to be those with parallel datasets of fewer than 1 million words. Most standard MT datasets contain tens or hundreds of millions of words.

[2] Note that in this analysis, we do not use the dictionaries, only complete sentences of training data.



(a) Tokens

(b) Types

Figure 6.1: Token-based and type-based OOV rates across six Indian languages. The curves are generated by randomly sampling the training datasets described in Section 6.4.1.

Source	গাণিতিকভাবে	ফাংশন	অভিষেক	পোষাকও	ফুটনোট	বোঝার
Induced Translations	mathematical equal ganitikovabe	**function** functions variables	made goal earned	shaky pashan shirts	mutant futbol futebol	vain newton boer
Correct Translation	mathematically	function	inauguration	dress	footnote	understand

Figure 6.2: Examples of OOV Bengali words, our top-3 ranked induced translations, and their correct translations. Correct induced translations are bolded.

As before, we use a diverse set of features estimated over comparable corpora and a small set of known translations as supervision for training a discriminative classifier, which makes predictions (translation or not a translation) on test set words paired with all possible translations. Possible translations are taken from the set of all target words appearing in the comparable corpora. Candidates are ranked according to their classification scores. In the settings that we explore in this chapter, we have access to either a seed bilingual dictionary or a small parallel corpus, which makes such a supervised approach to bilingual lexicon induction a natural choice.

We use the same feature set as in Chapter 4, which includes the temporal, contextual, topic, orthographic, and frequency similarity between a candidate translation pair. We derive translations to serve as positive supervision from our existing bilingual dictionaries or from automatically aligned parallel text[3] and, as before, use random word pairs as negative supervision, depending on the experimental setting. Figure 6.2 shows some examples of Bengali words, their correct translations, and the top-3 translations that this framework induces.

In our initial experiments, we add the single highest ranked English candidate translation for each source language OOV to our phrase tables. Adding these translations by definition improves the coverage of our MT models.

[3]GIZA++ intersection alignments over all training data.

6.2 Improving Accuracy

Following the phrase table scoring methods presented in Chapter 5, we use comparable corpora to estimate additional features over the translation pairs in our phrase tables and include those features in tuning and decoding in order to improve the accuracy of our models. We compute both phrasal and lexical features for all of the following except orthographic similarity, for which we only use lexically smoothed features, resulting in nine additional features: temporal similarity based on time-stamped web crawls, contextual similarity based on web crawls and Wikipedia (separately), orthographic similarity using normalized edit distance, and topic similarity based on inter-lingually linked Wikipedia pages. As before, we use time-stamped web crawl data to estimate temporal similarity, Wikipedia interlingual links to estimate topic similarity, and both corpora to estimate contextual similarity. Our expectation is that by adding a diverse set of similarity features to the phrase tables, our models will better distinguish between good and bad translation pairs, improving accuracy.

6.3 Zero Parallel Data Setting

In this section, we assume that no parallel corpus of translated sentences is available for training. Instead, we build statistical translation models using existing bilingual dictionaries, transliterations, and induced translations and translate several Wikipedia pages in each of 22 languages. We chose to translate the following source language Wikipedia pages because they are familiar topics, cover a variety of subjects, and exist in many languages: Barack Obama, Islam, and Forest.

We generate phrase tables for each source language based on (1) bilingual dictionaries, (2) transliterations, and (3) induced translations for OOV words. We described our bilingual dictionaries in Section 3.2.2. Our dictionaries consist mostly of unigram translations but do include some multiword phrase pairs. As we described in Section 4.2.1.3, we do transliteration by training character-based translation models on Wikipedia page titles. We generate the 1-best transliteration for all non-domain script source language words. We used the data, features, and reranking methods described in Chapter 4 to propose 10 translations for each OOV source language word. Then, we score each patchwork phrase table using the same similarity features described above: web crawls contextual similarity, web crawls temporal similarity, Wikipedia contextual similarity, Wikipedia topic similarity, and orthographic similarity. We compute both phrasal and lexical translation features. Additionally, we estimate a lexicalized reordering model from our Wikipedia comparable corpora using the algorithm that we proposed in Klementiev *et al.* (2012). We estimate a language model over the entire English Wikipedia corpus except those five topical pages which we wish to translate.

Typically in statistical machine translation, in addition to using parallel corpora to estimate the parameters of an SMT model, a small bitext is also used as a development set to tune the feature weights of the log linear model. For many of the low resource languages that we experiment with here, such data is not readily available even in the small quantities needed for tuning. Rather than gather tuning sets for each language, we reuse the weights that were learned for a Bengali-English MT experiment[4] that used the same set of monolingually derived features. Of course, the source language and corpora change substantially in these new experiments, and the optimal weights are unlikely to be the same. In the next section, 6, we make use of small amounts of bitext to seed a phrase table, tune feature weights, and also for automatic evaluation.

Because we do not have translations for source language Wikipedia pages, we cannot evaluate translation quality with an automatic metric like BLEU. However, because the topics are familiar, it is possible to read the output and get a qualitative sense of the translation quality. Table 14.1 in Appendix 14 shows the first few lines of each source language page on *Barack Obama* translated into English. In Section 6.4.3, we present a variety of BLEU score results on test sets for which we do have reference translations.

Figures 6.3 and 6.4 show the first few sentences of the Hindi Wikipedia pages on *Forest* and *Islam* translated several ways. In each figure, the Hindi source paragraph is given in (1) followed by a dictionary gloss (2) and a transliteration gloss (3). The dictionary glosses are based on our original bilingual dictionaries (see Table 11.2 in Appendix 11). If the dictionaries contain more than one translation of a given word, we pick one randomly. We transliterate each Hindi word to obtain the transliteration glosses. We use the machine translation based approach to transliteration presented in Irvine *et al.* (2010b) and, like that work, train models based on Wikipedia person-page titles. In both sets of translations, the dictionary glosses are somewhat readable, but there are many OOV words. The transliterations, in contrast, are not nearly as readable. Although the transliterations of some cognates, including *hayadrologik* and *biosphia* in the forest

[4]The Bengali model was chosen at random.

translation, are understandable, most words are not. In our experiments, we have seen that the number of cognates and named entities, which can often be accurately transliterated instead of translated, vary by subject matter. For example, in the Hindi page on Barack Obama, there are many more 'transliteratable' words than the Hindi page on forests.

The 'Dictionary + Transliterations + Monolingual Scoring' (4) translation model uses a phrase table of dictionary translations and transliterations (used for the two word glosses) and our approach to monolingually scoring a phrase table described in Chapter 5, including both translation and reordering scoring. The next translation (5) is the same as the previous but also includes top-10 translations that we induce for each Hindi word by the methods presented in Chapter 4. Translation 5 is equivalent to those presented in Section 6.3 and uses no parallel data for training. For translations 4-5, as well as 6-7, we use a 5-gram language model trained on the English gigaword corpus. Translations 4 and 5 both have the ability to combine the strengths of our dictionary translations and our transliterations, and both are much better translations than either gloss. Scoring the dictionary pairs and the transliterations monolingually allows our translation model to learn how to distinguish competing translations for a given word. Introducing induced translations has a few noticeably good effects. For example, in the first sentence of the forest translation, the transliterations *uchucha*, *esjangal*, and *podahe* are used in the 'Dictionary + Transliterations + Monolingual Scoring' model. However, the next translation uses the induced translations *systolic*, *canopy*, and *headless* instead of the transliterations. It's unlikely that any of these words is a completely accurate translation, but they are closer than the non-English transliterations.

Translation 6 is produced by the model trained on our small Hindi training bitext (used in Section 6.4) and the gigaword-based English language model. The final translation (7) takes advantage of our entire bag of tricks: the small training bitext, our bilingual dictionaries, transliterations, induced translations, and monolingual scoring. The phrase table is populated with the top-10 induced translations, top-1 transliterations, dictionary pairs, and phrase pairs extracted from the word aligned training text. Each phrase pair is scored monolingually and those taken from the bitext are also scored bilingually. We use bilingually-estimated reordering scores when they are available and monolingually estimated scores (Klementiev *et al.*, 2012) for the remaining phrase pairs. Like the dictionary word gloss, using the model trained on the small bitext to translate the Hindi text alone results in many OOV words. However, using the small bitext allows us to accurately translate common words and phrase and function words, for example *which is* and *one of the most important*. Qualitatively, we prefer the final translation for each Hindi paragraph, which takes advantage of both bilingual and monolingual resources. This is consistent with the BLEU score results presented in Section 6.4.3.

6.4 Small Parallel Corpora Setting

In the next set of experiments, we begin with a baseline SMT model learned from a small parallel corpus and augment the model to improve its coverage and accuracy. As before, we combine techniques that take advantage of a variety of signals that can be estimated from comparable corpora. As in Chapter 4, we estimate a variety of signals of translation equivalence and combine those signals in order to predict translations for OOV words. Additionally, as in Chapter 5, we use the same comparable corpora and signals to estimate translation feature scores for new and existing translation pairs in our SMT model. We see improvements in translation quality between 0.5 and 1.4 BLEU points translating the following low resource languages into English: Tamil, Telugu, Bengali, Malayalam, Hindi, and Urdu.[5]

6.4.1 Data

Post *et al.* (2012) used Mechanical Turk to collect small parallel corpora for the following Indian languages and English: Tamil, Telugu, Bengali, Malayalam, Hindi, and Urdu. They collected both parallel sentence pairs and a dictionary of word translations.[6] We use all six datasets, which provide real low resource data conditions for six truly low resource language pairs. Table 6.1 shows statistics about the datasets.

As usual, we use both our web crawls and our Wikipedia comparable corpora for each language pair. Dataset sizes are given in Table 11.3 in Appendix 11.

6.4.2 Experimental setup

We use the training/development/test data splits given by Post *et al.* (2012) and, following that work, include the dictionaries in the training data and report results on the devtest set using case-insensitive BLEU and four references.

[5] We published the results presented in this section in Irvine and Callison-Burch (2013a).
[6] No dictionary was provided for Hindi.

Original Text (1)

एक वन एक उच्च घनत्व के साथ एक क्षेत्र है पेड़ -lrb- tree -rrb- एसजंगल के कई परिभाषाएँ , है जो की विभिन्न मानदंडों पर आधारित हैं.यह पोदाहे लगभग ९.४ % पृथ्वी की सतह को घेर ते है -lrb- या 30 % -rrb- जो की आवासों -lrb- habitat -rrb- ह्यद्रोलोगिक प्रवाह -lrb- hydrologic flow -rrb- मोदुलातोर्स -lrb- modulator -rrb- , और मिट्टी -lrb- soil -rrb- बचाव , एक पृथ्वी के बीओस्फिअ का सर्वाधिक महत्वपूर्ण पहलुओं के गठन.का प्रवास करते है इतिहास बताता है , की " वन " एक बीहड़ क्षेत्र जिसका मतलब कानूनी तौर पर बाजू के लिए निर्धारित शिकार -lrb- hunting -rrb- के द्वारा सामंती -lrb- feudal -rrb- कुलीनता -lrb- nobility -rrb- है , और इन शिकार जंगलों जरूरी ज्यादा अगर में सभी -lrb- देखें जंगली नहीं थे रॉयल वन -lrb- royal forest -rrb- -rrb- .हालांकि , शिकार के जंगलों अक्सर वुडलैंड के महत्वपूर्ण क्षेत्रों को शामिल किया जबकि , शब्द वन अंततः जंगली भूमि अधिक सामान्यतः मतलब करने के लिए आया था.एक वुडलैंड -lrb- woodland -rrb- जो की एक जंगल से भिन्न है .

Dictionary Word Glosses (2)

one forest one उच्च density its साथ one field is wood (tree) एसजंगल its lots definitions , is which of various मानदंडों on based है.यह पोदाहे total ९.४ % the earth of surface को surround ते is (either 30 %) those of आवासों (habitat) ह्यद्रोलोगिक प्रवाह (hydrologic flow) मोदुलातोर्स (modulator) , and soil (soil) safeguard , one the earth its बीओस्फिअ का rules important sides of गठन.का foreign do is history telling is , of " forest " one बीहड़ field whose means कानूनी for on बाजू of for nidhirit शिकार (hunting) its द्वारा सामंती (feudal) कुलीनता (nobility) is , and these शिकार in jungles compulsory more if me all (see wild no was royal forest (royal forest)) .हालांकि , शिकार its in jungles usual वुडलैंड its importance areas को शामिल did while , शब्द forest at the end wild land more generally means do of for was था.एक वुडलैंड (woodland) which of one जंगल from different is .

Transliteration Gloss (3)

ak vn ak uchcha ghantwa ke sath ak ksatra ha ped (tree) esjangal ke ki pribhashaën , ha jo ki vibhinn mandndon pr adharit han.yh podahe lgbhag . % prithvi ki sath ko gher te ha (ya 30 %) jo ki avason (habitat) hayadrologik prawah (hydrologic flow) modulators (modulator) , mitti (soil) bchaw , ak prithvi ke biosphia ka sarveadhik mahatwpurn phluon ke gthn.ka prawas krate ha dharampal battata ha , ki " vn " ak bihd ksatra jiska mtalb kanuni taur pr baju ke lier nirdharit shikar (hunting) ke dwara samanti (feudal) kulenta (nobility) ha , in shikar junglon jruri jayada agar man sbhi (dekhen jungali nhin the royle vn (royal forest)) .hallanki , shikar ke junglon aksr woodland ke mahatwpurn ksatron ko shamill kiya jbki , shbd vn antt: jungali bhumi adhik samanyat: mtalb karne ke lier aya tho.aq woodland (woodland) jo ki ak jangal se bhinn ha .

Dictionary + Transliterations + Monolingual Scoring (4)

one forest one uchcha density of sath one field is tree (tree) esjangal of many definitions , is which of various mandndon on based han.yh podahe nearly . % of the earth surface ko surround te is (or 30 % of which) avason (habitat) hayadrologik prawah (hydrologic flow) modulators (modulator) , and soil (soil) safeguard , one of the earth biosphia ka more important sides of gthn.ka foreign to do is history telling is , of " forest " one bihd field whose means kanuni for on its baju for nidhirit shikar (hunting) of dwara samanti (feudal) kulenta (nobility) is , and these shikar forests necessary more if among all (see no wild was royal forest (royal forest)) .hallanki , shikar of forests often woodland of important areas ko shamill did while , shbd forest at the end wild land more generally means do its for was tho.aq woodland (woodland) which of one jangal from different is .

Dictionary + Transliterations + Induced + Monolingual Scoring (5)

one forest one systolic density of which one field is tree (tree) canopy of many definitions , is which of various crm on based han.yh nearly headless . % of the earth surface ko surround te is (or 30 %) which of keyhole (organisms) canopy irr (telecom low) modulators (coniferous) , and soil (erosion) safeguard , one the earth of app ka more important sides of gthn.ka foreign to do is history telling is , of " forest " one maestra field whose means responsibility for on pulleys of for nidhirit mane (africana) of dhara necker (electors) émigrés (forest) is , and these lions forests more necessary if among all (see no wild the royal forest (royal society)) .hallanki , mane of forests often evergreen of important areas ko they did while , quirk forest at the end wild land more generally means do its for was tho.aq evergreen (forests) which of one forest from different is .

Small Bitext Translation (6)

a forest with a high density is one area of the tree (एसजगल) many definitions of the tree , which is full of various मानदंडों पोदाहे based on almost ९.४ % to the surface of the earth is ' घेर) (or 30 % of the habitat , which produced (flow)) (flow ह्यद्रोलोगिक (hydrologic) मोदुलातोर्स (modulator) soil , and the will of the earth , one of the most important scripts बीओस्फिअ गठन.का . history tells everybody , " " the " " forest , which a बीहड़ area on legal means for बाजू (hunting victim decided by the feudal) कुलीनता ((सामंती)) nobility , and these victim is more important जंगलों (, forests , see if all in were not royal forest .हालांकि ( )) royal forest of the victim , often वुडलैंड जंगलों included to the important areas , while in the forest , word means more generally , forest land for वुडलैंड came to woodland ( ) था.एक , which is different from the one from .

Small Bitext Trans + Dict + Translit + Induced + Mono Scoring (7)

a high density of a forest area is with a tree (tree) canopy of definitions , which is one of the many different crm based on this canopy almost . % of the earth 's surface to surround ' is (or 30 %) which houses , (elephants) canopy flow (hydrologic low) canopy (canopy) , and the soil (erosion) saved , one of the most important earth 's monoculture gthn.ka journey of scripts . " " he tells the history , " " a forest area maestra , whose legal means for the pulley on decided victim (elephants) by tong (electors) danton (forest) , and more important forests victim if these all in (see was not , wild royal forest (royal forest)) .hallanki , victim of the evergreen often forests of important areas , while . finally , wild forest land , the word means more commonly used to refer to came tho.aq evergreen (forests) , which is different from the one from .

Figure 6.3: First paragraph of Hindi Wikipedia page on *Forest*, and a progression of translations of it.

Original Text (1)

इस्लाम धर्म -lrb- الإسلام -rrb- ईसाई धर्म के बाद अनुयाइयों के आधार पर दुनिया का दूसरा सब से बड़ा धर्म है । इस्लाम शब्द अरबी भाषा का शब्द है जिसका मूल शब्द सल्लमा है जिस की दो परिभाषाएं हैं -lrb- १ -rrb- अमन और शांति -lrb- २ -rrb- आत्मसमर्पण ।

ईस्लाम एकेश्वरवाद को मानता है । इसके अनुयायियों का प्रमुख विश्वास है कि ईश्वर सिर्फ़ एक है और पूरी सृष्टि में सिर्फ़ वह ही महिमा -lrb- इबादत -rrb- के लायक है , और सृष्टि में हर चीज़ , ज़िंदा और बेजान , दृश्य और अदृश्य उसकी इच्छा के सामने आत्मसमर्पित और शांत है । इस्लाम धर्म की पवित्र पुस्तक का नाम क़ुरआन है जिसका हिंदी में मतलब सस्वर पाठ है । इसके अनुयायियों को अरबी में मुस्लिम कहा जाता है , जिसका बहुवचन मुसलमान होता है । मुसलमान यह विश्वास रखते है कि क़ुरआन जिब्राईल -lrb- ईसाईयत में gabriel -rrb- नामक एक फ़रिश्ते के द्वारा , मुहम्मद साहब को ७वीं सदी के अरब में , लगभग २३ साल में याद-कंठस् थ कराया गया था । मुसलमान इस्लाम को कोई नया धर्म नहीं मानते । उनके अनुसार ईश्वर ने मुहम्मद साहब से पहले भी धरती पर कई दूत भेजे हैं , जिनमें इब्राहीम , मूसा और ईसा शामिल हैं । मुसलमानों के अनुसार मूसा और ईसा के कई उपदेशों को लोगों ने विकृत कर दिया । अधिकतम मुसलमानों के लिये मुहम्मद साहब ईश्वर के अन्तिम दूत थे और क़ुरआन मनुष्य जाति के लिये अन्तिम संदेश है ।

Dictionary Word Glosses (2)

islam religion (الإسلام) christian religion of after अनुयाइयों its foundation / support on world का another all from huge religion is । islam शब्द arabic language का शब्द is whose worth शब्द सल्लमा is from which of दो परिभाषाएं are (१) अमन and peace (२) dedicated ।

ईस्लाम एकेश्वरवाद को believe is । for this followers का major विश्वास is that god सिर्फ़ one is and complete universe me सिर्फ़ that only महिमा (worship) its लायक is , and universe me every thing , ज़िंदा and बेजान , view and invisible his desire its front आत्मसमर्पित and quiet is । islam religion of holy book का name kuran -holy book of islam is whose hindi me means सस्वर lesson is । for this followers को arabic me मुस्लिम कहा go is , whose बहुवचन muslim happens is । muslim यह विश्वास रखते is that koran जिब्राईल (ईसाईयत among gabriel) नामक one फ़रिश्ते of द्वारा , mohammad saheb को ७वीं century its arab me , total 23 year me याद-कंठस् tha कराया went था । muslim islam को कोई नया religion no know as । his अनुसार god ने mohammad saheb from before also earth on lots embassdor sent are , in which इब्राहीम , musa and ईसा शामिल is । muslims its अनुसार musa and ईसा of many lectures को people ने विकृत do gave । maximum muslims its for mohammad saheb god of अन्तिम embassdor was and koran मनुष्य race of for अन्तिम message is ।

Transliteration Gloss (3)

islam dharam (الإسلام) isai dharam ke bad anuyaiyon ke adhar pr dunia ka dusara sb se bdha dharam ha . islam shbd arbi bhasha ka shbd ha jiska moole shbd sallama ha gis ki do pribhashaën han (i) aman shanti (ii) atamsamarpn .

islam akeshwarvad ko manta ha . iske anuyayion ka pramukh viswaas ha ki ishwar sierf ak ha puri s.ti man sierf vh hi mahima (ibadt) ke laik ha , s.ti man har chez , zinda began , drishy adrishy usky iachha ke samane atamsamarpit shant ha . islam dharam ki pavitra pustek ka nam quran ha jiska hindi man mtalb sswar path ha . iske anuyayion ko arbi man muslim kha jata ha , jiska bahuvchn musalman hota ha . musalman yh viswaas rkhte ha ki quran gibrail (isaiyat man gabriel) namk ak frishte ke dwara , muhammad sahb ko viiwin sdi ke arb man , lgbhag ii sal man yad-kanthus th karaya ghiya tha . musalman islam ko koi nya dharam nhin manate . unce anussaur ishwar ne muhammad sahb se phle bhi dhrati pr ki dut bhage han , jinman ibrahim , musa isa shamill han . musalmanon ke anussaur musa isa ke ki upadeshon ko logon ne vicrit kar dia . adhictam musalmanon ke lier muhammad sahb ishwar ke anthim dut the quran manushy jati ke lier anthim sandesh ha .

Dictionary + Transliterations + Monolingual Scoring (4)

islam religion (الإسلام) christian religion of after anuyaiyon its basis on world ka second all from big religion is . islam shbd arabic language ka shbd is whose original shbd sallama is from which of do pribhashaën is (1) aman and peace (ii) dedicated .

islam akeshwarvad ko believe is . its followers ka major viswaas is that god sierf one is and complete universe among sierf that only mahima (worship) of laik is , and universe among each and every thing zinda , and began , view and invisible his desire of front atamsamarpit and quiet is . islam religion of holy book ka name quran is whose hindi among means sswar path is . its followers ko arabic among muslim kha go is , whose bahuvchn muslim happens is . muslim yh viswaas rkhte is that koran gibrail (isaiyat among gabriel) namk one frishte of dwara , muhammad saheb ko viiwin century of arab man , nearly 23 year among yad-kanthus tha karaya made tha . muslim islam ko koi nya no religion feel . his anussaur god ne muhammad saheb from before also earth on many dut is sent , in which ibrahim , musa and isa shamill is . muslims of anussaur musa and isa of many lectures ko people ne vicrit do give . maximum muslims of for muhammad saheb god of anthim dut and the koran manushy race for its anthim message is .

Dictionary + Transliterations + Induced + Monolingual Scoring (5)

islam religion (alevi) christian religion of after adulation of basis on world ka second all from big religion is . islam quirk arabic language ka quirk is whose original quirk isis is from which of do rima is (1) aman and peace (ii) dedicated .

islam anic ko believe is . its followers ka major undead is that god sirf one is and complete universe among sirf that only pardes (worship) of below is , and universe among each and every thing , sexiest and began , view and invisible his desire of front ndf and quiet is . islam religion of holy book ka name koran is whose hindi among means guttural path is . its followers ko arabic among muslim who go is , whose verbs muslim happens is . muslim yh undead there is that quran reciters (crucifixion among pen) took one of frisbee dhara , muhammad saheb ko nagari century of arab man , nearly 23 year among yad-kanthus tha online made tha . muslim islam ko but this no religion feel . his like god ne muhammad saheb from before also earth on many dut is sent , in which suras , genesis and middle there is . muslims of being genesis and middle of many lectures ko people ne folklore do had . maximum muslims of for muhammad saheb god of shunga dut and the koran manus race of for shunga message is .

Small Bitext Translation (6)

islam religion ( ) الإسلام after the christian religion on the basis of followers to the world 's second largest religion . islam is the word of the arabic language is the word , which is the word salma of which are two means अमन (1) (2) submission and peace .

ईस्लाम believes in monotheism . its followers believe that god is the only one in the whole universe , and only the glory of the (लायक ,) , and all the things in the creation and बेजान चीज़ , ज़िंदा , god , are at his will and quiet . islam 's holy book is the name of the quran , which means recitation in hindi . its followers are called muslims in arabic , is the plural of muslims . muslims believe that quran in christianity) gabriel (gabriel , by the faristha named to prophet muhammad in the 7th century in arabia , almost 23 years in the verses of the याद-कंठस् was made to islam . muslims do not believe that there is no new religion . according to him , before the prophet muhammad , god has also sent many messengers on earth , in which includes , musa and isa are included . according to the muslims , musa and isa many people pervert . for the majority of muslims mohammed saheb was the last prophet of god and quran for the human race is the last message .

Small Bitext Trans + Dict + Translit + Induced + Mono Scoring (7)

islam , christianity) alevi (religion after the followers on the basis of the world is the second largest religion islam . word of the arabic language from which means is are (1) aman and peace (2) surrender .

islam monotheism . this is to the head of the followers believe that there is only one god , and only in the whole world , he is the lack of) worship (, and in every thing , sexiest creation and imran , god , in front of the wish and are peaceful . islam 's holy book is the name of the quran recitation in hindi , which means its adherents . muslims in arabic , it is said , whose big muslim . muslims believe that quran gabriel (christianity in gabriel) named a frisbee by the , to the prophet muhammad in the 7th century , the arab almost 23 years in the verses of the yad-kanthus . muslims , islam is a new religion . do not believe according to mohammed before god has also sent envoy on earth , in which there are many including , musa bc . muslims are included , and according to the teachings of musa bc , and many people to deform . many muslims for the messenger of god 's prophet muhammad was the last of the quran and the last message for mankind .

Figure 6.4: First paragraph of Hindi Wikipedia page on *Islam*, and a progression of translations of it.

Language	Words of Training Data		Dev Types	Dev Tokens
	Sentences	Dictionary	% OOV	% OOV
Tamil	334,714	77,240	44	25
Telugu	414,094	40,742	39	21
Bengali	239,555	6,783	37	18
Malayalam	263,086	151,194	6	3
Hindi	658,977	0	34	11
Urdu	615,635	116,496	23	6

Table 6.1: Information about datasets released by Post *et al.* (2012): words in the source language parallel sentences and dictionaries, and percent of development set word types and tokens that are OOV (do not appear in either section of the training data).

Language	Top-1 Acc.	Top-10 Acc.
Tamil	4.5	10.2
Telugu	32.8	47.9
Bengali	17.9	29.8
Malayalam	12.9	23.0
Hindi	44.3	57.6
Urdu	16.1	33.8

Table 6.2: Percent of word types in a held out portion of the training data which are translated correctly by our bilingual lexicon induction technique. Evaluation is over the top-1 and top-10 outputs in the ranked lists for each source word.

We use the Moses phrase-based MT framework (Koehn *et al.*, 2007). For each language, we extract a phrase table with a phrase limit of seven. In order to make our results comparable to those presented in Post *et al.* (2012), we follow that work and use the English side of the training data to train a language model. Using a language model trained on a larger corpus (e.g. the English side of our comparable corpora) may yield better results, but such an improvement is orthogonal to the focus of this work. Throughout our experiments, we use the batch version of MIRA (Cherry and Foster, 2012) for tuning the feature set.[7] We rerun tuning for all experimental conditions and report results averaged over three tuning runs (Clark *et al.*, 2011).

Our baseline uses the bilingually extracted phrase pairs and standard translation probability features. We augment it with the single top ranked translation for each OOV to improve coverage (+ OOV Trans) and with additional features to improve accuracy (+Features). We make each modification separately and then together. Then we present additional experiments where we induce translations for low frequency words, in addition to OOVs (6.4.3.3), append top-k translations (6.4.3.4), vary the amount of training data used to induce the baseline model (6.4.3.5), and vary the amount of comparable corpora used to estimate features and induce translations (6.4.3.6).

6.4.3 Results

6.4.3.1 Bilingual Lexicon Induction

Before presenting end-to-end MT results, we examine the performance of the supervised bilingual lexicon induction technique that we use for translating OOVs. In Table 6.2, top-1 accuracy is the percent of source language words in a held out portion of the training data [8] for which the highest ranked English candidate is a correct translation. Post *et al.* (2012) gathered up to six translations for each source word, so some have multiple correct translations. Performance is lowest for Tamil and highest for Hindi. For all languages, top-10 accuracy is much higher than the top-1 accuracy. In Section 6.4.3.4, we explore appending the top-k translations for OOV words to our model instead of just the top-1.

[7] We experimented with MERT and PRO as well but saw consistently better baseline performance using batch MIRA.
[8] We retrain with all training data for MT experiments.

Experiment	Tamil		Telugu		Bengali		Malayalam		Hindi		Urdu	
	BLEU	Diff.	BLEU	Diff.	BLEU	Diff.	BLEU	Diff.	BLEU	Diff.	BLEU	Diff.
Baseline	9.45		11.72		12.07		13.55		15.01		20.39	
+Features	9.77	+0.32	11.96	+0.24	12.25	+0.18	14.15	+0.60	15.34	+0.33	20.97	+0.58
+OOV Trans.	9.45	0.00	12.20	+0.48	**12.74**	+0.67	13.65	+0.10	15.59	+0.58	21.30	+0.91
+Feats & OOV	**9.98**	+0.53	**12.25**	+0.53	12.55	+0.48	**14.18**	+0.63	**16.08**	+1.07	**21.78**	+1.39
OOV Oracle	12.32	+2.87	16.04	+4.32	16.41	+4.34	13.55	0.00	17.72	+2.71	22.80	2.41
Hiero	9.81		12.46		12.72		13.72		15.53		19.53	
SAMT	9.85		12.61		13.53		14.28		17.29		20.99	

Table 6.3: BLEU performance gains that target coverage (+OOV Trans.) and accuracy (+Features), and both (+Feats & OOV). OOV oracle uses OOV translations from automatic word alignments. Hiero and SAMT results are reported in Post *et al.* (2012).

6.4.3.2 Improving Coverage and Accuracy in End-to-End SMT

Table 6.3 shows our results adding OOV translations, adding features, and then both. Additional translation features alone, which improve our models' accuracy, increase BLEU scores between 0.18 (Bengali) and 0.60 (Malayalam) points.

Adding OOV translations makes a big difference for some languages, such as Bengali and Urdu, and almost no difference for others, like Malayalam and Tamil. The OOV rate (Table 6.1) is low in the Malayalam dataset and high in the Tamil dataset. However, as Table 6.2 shows, the translation induction accuracy is low for both. Since few of the supplemental translations are correct, we don't observe BLEU gains. In contrast, induction accuracies for the other languages are higher, OOV rates are substantial, and we do observe moderate BLEU improvements by supplementing phrase tables with OOV translations.

In order to compute the *potential* BLEU gains that we could realize by correctly translating all OOV words (achieving 100% accuracy in Table 6.2), we perform an oracle experiment. We use automatic word alignments over the test sets to identify correct translations and append those to the phrase tables.[9] The results, in Table 6.3, show possible gains between 4.3 (Telugu and Bengali) and 0 (Malayalam) BLEU points above the baseline. Not surprisingly, the possible gain for Malayalam, which has a very low OOV rate, is very low. Our *+OOV Trans.* model gains between 0% (Tamil) and 38% (Urdu) of the potential improvement.

Using comparable corpora to improve both accuracy (+Features) and coverage (+OOV Trans.) results in translations that are better than applying either technique alone for five of the six languages. BLEU gains range from 0.48 (Bengali) to 1.39 (Urdu). We attribute the particularly good Urdu performance to the relatively large comparable corpora (Table 11.3). In Section 6.4.3.6, we present results varying the amount of Urdu-English comparable corpora used to induce translations and estimate additional features.

Table 6.3 also shows the Hiero (Chiang, 2005) and SAMT (Zollmann and Venugopal, 2006) results that Post *et al.* (2012) report for the same datasets. Both syntax-based models outperform the phrase-based MT baseline for each language except Urdu, where the phrase-based model outperforms Hiero. Here, we extend a phrase-based rather than a syntax-based system because it is simpler. However, we expect that our improvements will also apply to syntactic models. Because our efforts have focused on the accuracy and coverage of translation pairs and have not addressed reordering or syntax, we expect that combining them with an SAMT grammar will result in state-of-the art performance.

6.4.3.3 Translations of Low Frequency Words

Given the positive results in Section 6.4.3.2, we hypothesize that mining translations for low frequency words, in addition to OOV words, may improve accuracy. For source words which only appear a few times in the parallel training text, the bilingually extracted translations in the standard phrase table are likely to be inaccurate and incomplete. Augmenting a model with additional translations for low frequency words may fix some previously SENSE errors, in which a source word was observed in training but not with its correct translation. Therefore, we perform additional experiments varying the minimum source word training data frequency for which we induce additional translations. That is, if $freq(w_{src}) \leqslant M$, we induce a new translation for it and include that translation in our phrase table.

[9] Because the automatic word alignments are noisy, this oracle is conservative.

Language	Baseline	M: trans added for $freq(w_{src}) \leqslant M$					
		0	1	5	10	25	50
Tamil	9.5	10.0	9.9	**10.2**	**10.2**	9.9	**10.2**
Telugu	11.7	12.3	12.2	12.3	**12.4**	12.3	11.9
Bengali	12.1	12.6	12.8	13.0	12.9	**13.1**	13.0
Malayalam	13.6	**14.2**	14.1	**14.2**	**14.2**	13.9	13.9
Hindi	15.0	16.1	16.1	**16.2**	**16.2**	16.0	15.8
Urdu	20.4	21.8	21.8	21.8	21.9	**22.1**	21.8

Table 6.4: Varying minimum parallel training data frequency of source words for which new translations are induced and included in the phrase-based model. In all cases, the top-1 induced translation is added to the phrase table and features estimated over comparable corpora are included (i.e. +Feats & Trans model).

Note that in the results presented in Table 6.3, $M = 0$. In these experiments, we include our additional phrase table features estimated over comparable corpora and hope that these scores will assist the model in choosing among multiple translation options for low frequency words, one or more of which is extracted bilingually and one of which is induced using comparable corpora. Table 6.4 shows the results when we vary M. As before, we average BLEU scores over three tuning runs.

In general, modest BLEU score gains are made as we augment our phrase-based models with induced translations of low frequency words. The highest performance is achieved when M is between 5 and 50, depending on language. The largest gains are 0.5 and 0.3 BLEU points for Bengali and Urdu, respectively, at $M = 25$. This is not surprising; we also saw the largest relative gains for those two languages when we added OOV translations to our baseline model. With the addition of low frequency translations, our highest performing Urdu model achieves a BLEU score that is 1.7 points higher than the baseline.

In different data conditions, inducing translations for low frequency words may result in better or worse performance. For example, the size of the training set impacts the quality of automatic word alignments, which in turn impacts the reliability of translations of low frequency words. However, the experiments detailed here suggest that including induced translations of low frequency words will not hurt performance and may improve it.

6.4.3.4 Appending Top-K Translations

So far we have only added the top-1 induced translation for OOV and low frequency source words to our phrase-based model. However, the bilingual lexicon induction results in Table 6.2 show that accuracies in the top-10 ranked translations are, on average, nearly twice the top-1 accuracies. Here, we explore adding the top-k induced translations. We hope that our additional phrase table features estimated over comparable corpora will enable the decoder to correctly choose between the *k* translation options. We induce translations for OOV words only ($M = 0$) and include all comparable corpora features.

Table 6.5 shows performance as we append the top-k ranked translations for each OOV word and vary k. With the exception of Bengali, using a k greater than 1 does not increase performance. In the case of Bengali, and additional 0.2 BLEU is observed when the top-25 translations are appended. In contrast, we see performance decrease substantially for other languages (0.7 BLEU for Telugu and 0.2 for Urdu) when the top-25 translations are used. Therefore, we conclude that, in general, the models do not sufficiently distinguish good from bad translations when we append more than just the top-1. Although using a k greater than 1 means that more correct translations are in the phrase table, it also increases the number of possible outputs over which the decoder must search.

6.4.3.5 Learning Curves over Parallel Data

In the experiments above, we only evaluated our methods for improving the accuracy and coverage of models trained on small amounts of bitext using the full parallel training corpora released by Post *et al.* (2012). Here, we apply the same techniques but vary the amount of parallel data in order to generate learning curves. Figure 6.5 shows learning cures for all six languages. In all cases, results are averaged over three tuning runs. We sample both parallel sentences and dictionary entries.

Language	Baseline	*k*: top-*k* translations added				
		1	3	5	10	25
Tamil	9.5	**10.0**	**10.0**	9.8	**10.0**	**10.0**
Telugu	11.7	**12.3**	11.7	11.9	11.7	11.6
Bengali	12.1	12.6	12.6	12.6	12.7	**12.8**
Malayalam	13.6	**14.2**	**14.2**	**14.2**	**14.2**	14.1
Hindi	15.0	**16.1**	16.0	15.9	15.9	15.9
Urdu	20.4	**21.8**	**21.8**	21.7	21.5	21.6

Table 6.5: Adding top-k induced translations for source language OOV words, varying *k*. Features estimated over comparable corpora are included (i.e. +Feats & Trans model). The highest BLEU score for each language is highlighted. In many cases differences are less than 0.1 BLEU.

All six learning curves show similar trends. In all experimental conditions, BLEU performance increases approximately linearly with the log of the amount of training data. Additionally, supplementing the baseline with OOV translations improves performance more than supplementing the baseline with additional phrase table scores based on comparable corpora. However, in most cases, supplementing the baseline with both translations and features improves performance more than either alone. Performance gains are greatest when very little training data is used. The Urdu learning curve shows the most gains as well as the cleanest trends across training data amounts. As before, we attribute this to the relatively large comparable corpora available for Urdu.

6.4.3.6 Learning Curves over Comparable Corpora

In our final experiment, we consider the effect of the amount of *comparable corpora* that we use to estimate features and induce translations. We present learning curves for Urdu-English because we have the largest amount of comparable corpora for that pair. We use the full amount of parallel data to train a baseline model, and then we randomly sample varying amounts of our Urdu-English comparable corpora. Sampling is done separately for the web crawl and Wikipedia comparable corpora. Figure 6.6 shows the results. As before, results are averaged over three tuning runs.

The phrase table features estimated over comparable corpora improve end-to-end MT performance more with increasing amounts of comparable corpora. In contrast, the amount of comparable corpora used to induce OOV translations does not impact the performance of the resulting MT system as much. The difference may be due to the fact that data sparsity is always more of an issue when estimating features over *phrase pairs* than when estimating features over *word pairs* because phrases appear less frequently than words in monolingual corpora. Our comparable corpora features are estimated over phrase pairs while translations are only induced for OOV words, not phrases. So, it makes sense that the former would benefit more from larger comparable corpora.

6.4.4 Post-Augmentation WADE Analysis

Earlier in the thesis, in Section 3.4.2, we presented a WADE analysis showing the relative frequency of SEEN, SENSE, and SCORE errors made by models trained on small amounts of bitext. In Section 6.4.3.2, we showed improvements in translation quality as measured by BLEU when we applied our methods for using comparable corpora to improve SEEN and SCORE errors. Here, we analyze our augmented models using WADE in order to better understand the effects of our comparable corpora-based modifications. We begin by presenting a version of WADE that uses multiple reference translations (Section 6.4.4.1) and then present the results of the analysis (Section 6.4.4.2).

6.4.4.1 WADE with Multiple References

Recall that the basic unit of analysis for WADE is an alignment link. With WADE, we word align (either manually or automatically) each source language test set sentence with its target language reference translation. Then, we compare the set of word alignments between a given source language test sentence and its machine translation with the reference links. Alternative machine translations are compared with respect to how many of the reference links they cover. Figure 6.7 shows an example. It is easy to compare the first sentence to the second using WADE because they

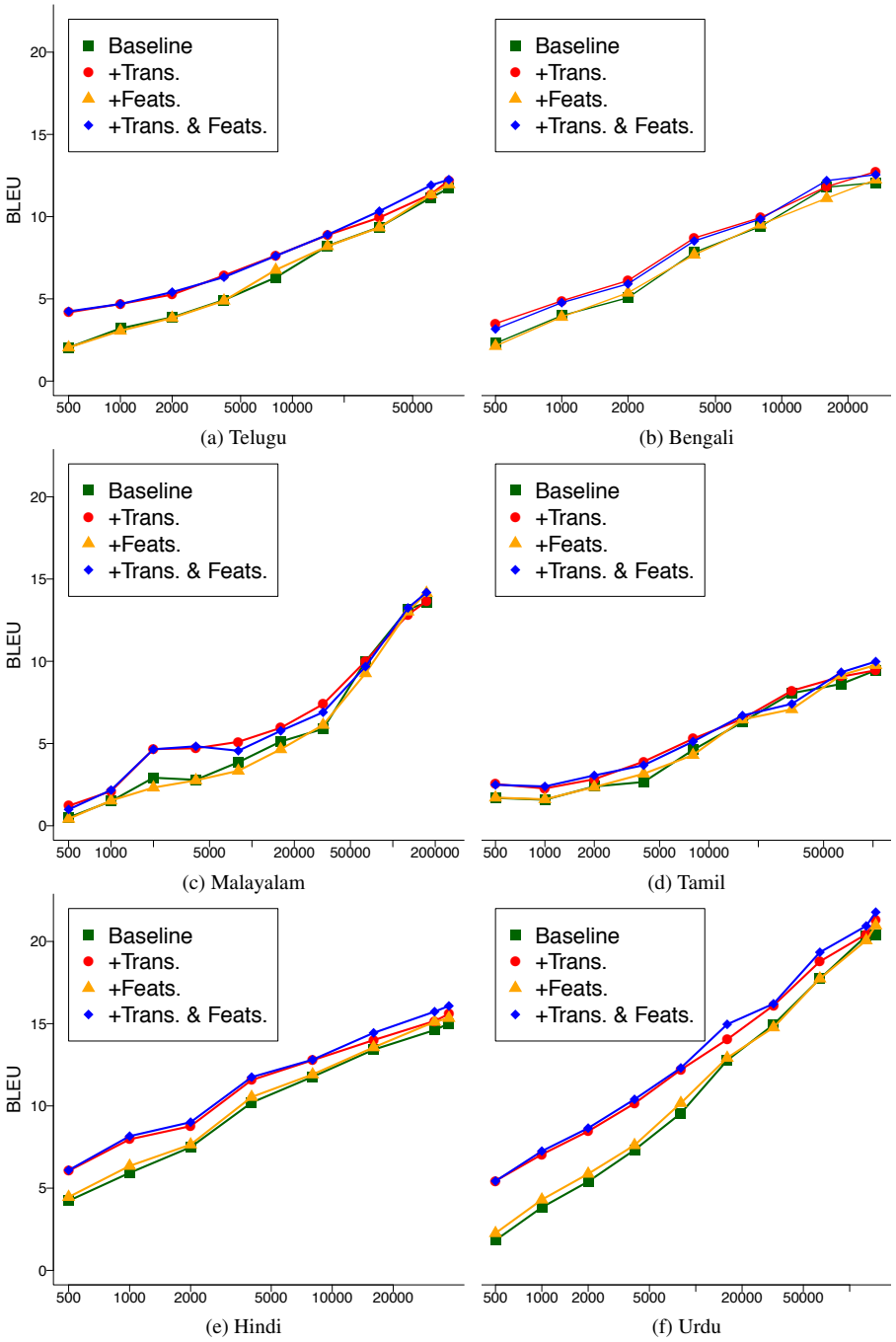


Figure 6.5: Comparison of learning curves over lines of parallel training data for four SMT systems: our baseline phrase-based model (baseline), model that supplements the baseline with translations of OOV words induced using our supervised bilingual lexicon induction framework (+Trans), model that supplements the baseline with additional phrase table features estimated over comparable corpora (+Feats), and a system that supplements the baseline with both OOV translations and additional features (+Trans & Feats).

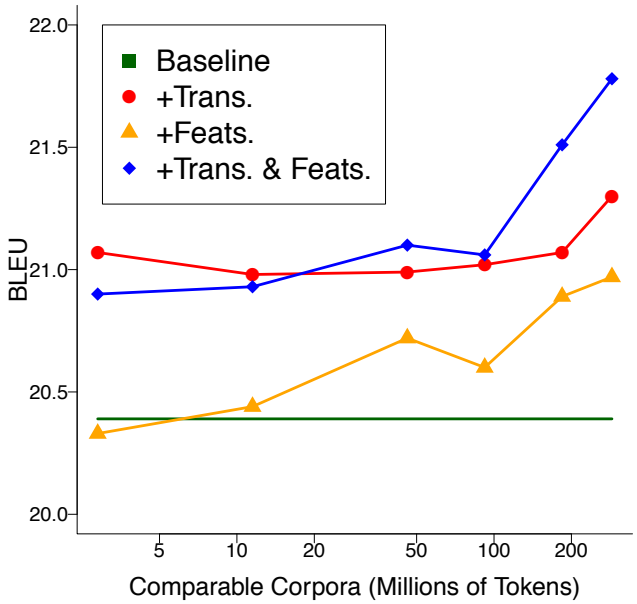


Figure 6.6: Urdu to English translation results using varying amounts of comparable corpora to estimate features and induce translations.

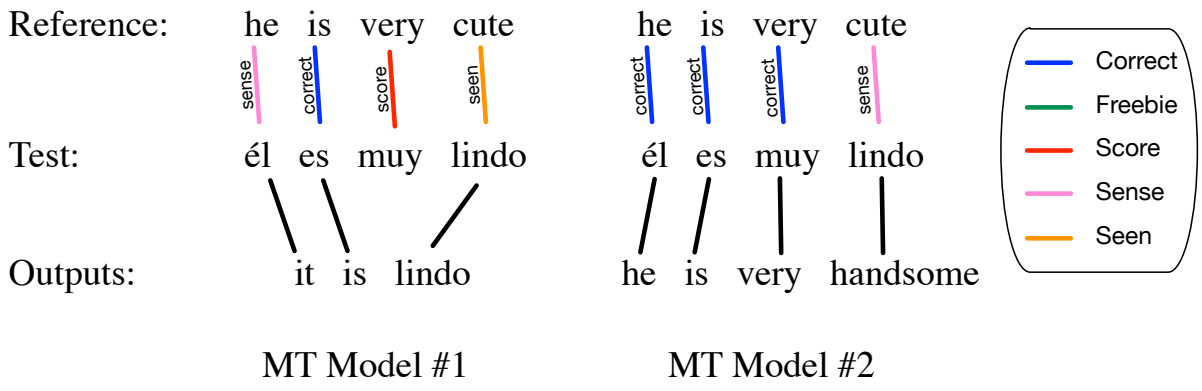


Figure 6.7: Example using WADE to compare two machine translations. Moving from the first machine translation to the second, one sense error is corrected, one score error is corrected, and one seen error becomes a sense error.

are each evaluated over the same set of four reference links. In the example, two errors in the first machine translation are corrected in the second, and another shifts from a SEEN error to a SENSE error.

In contrast to WADE, the basic unit of analysis in the BLEU metric is each ngram in the target language machine translation output. Generalizing the BLEU metric from one reference to many involves allowing each ngram in the output to match an ngram in any of the available reference translations. For example, if we had a second reference translation for the example given in Figure 6.7 that read *it is handsome*, then the second machine translation, *he is very handsome*, would achieve a perfect unigram precision even though neither reference contains all of the same word tokens.[10] The BLEU metric is intentionally simple, allowing for flexibility in the number of reference translations (Papineni *et al.*, 2002).

Because WADE's unit of analysis involves the reference translation itself, it is more difficult to generalize it to

[10]The lack of higher order ngram matches would prevent the output from receiving a perfect BLEU score.

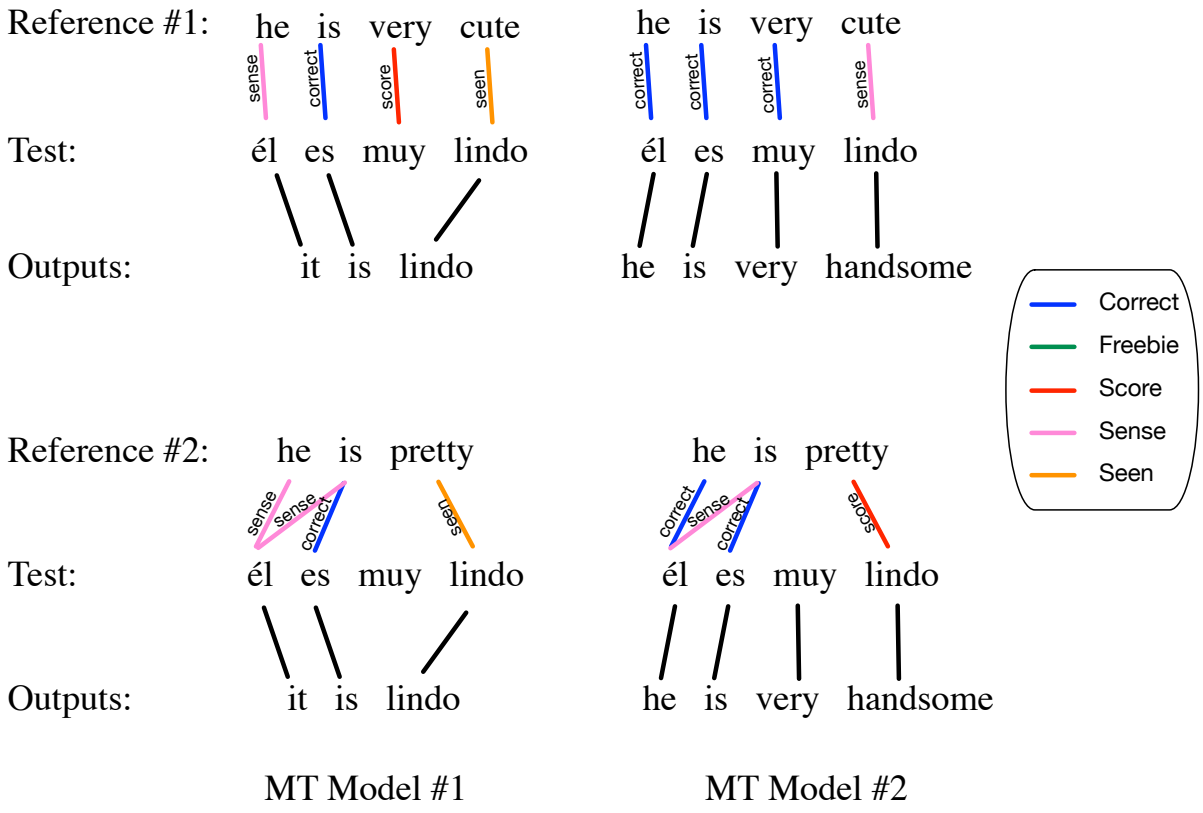


Figure 6.8: Example of the complications involved in using WADE to compare two machine translations using two reference translations. Identifying corresponding alignment links, comparing error categories, and dealing with incorrect automatic alignments are all challenges in doing a WADE analysis using multiple reference translations.

multiple references than it is to generalize BLEU. That is, BLEU measures the following: *is the output similar to the set of reference translations?* In contrast, WADE measures *is the reference translation included in the output and, if not, why?* BLEU measures precision over the machine translation output; WADE measures recall over the reference translations (and, specifically, over reference alignments, allowing for detailed error annotation). Generalization is further complicated by the fact that, in practice, we often use automatic word alignments and do no enforce consistency across partially overlapping reference translations.

Consider Figure 6.8. In comparison with Figure 6.7, we now have a second reference translation, *he is pretty*, and the goal is now to combine the analysis using the first (top) reference and the second (bottom) reference translation, with the ultimate goal of comparing the first (left) machine translation output with the second (right). In some cases, it is clear which reference alignment links correspond and how to combine them. For example, in the first reference *lindo* is aligned with *cute* and in the second reference *lindo* is aligned with *pretty*. The first MT model does not include a translation of *lindo*, so both reference alignments are marked as SEEN errors. It is clear that in any model of integrating the two reference translations to produce a single analysis, the first MT model should be credited with having made a SEEN error due to the OOV word *lindo*. The second MT model makes a SENSE error with respect to the first reference translation because it had not seen *lindo* translate as *cute*. However, it makes a SCORE error with respect to the second reference because, although the reference translation *pretty* was available in the model, it instead preferred *handsome*. In this case, it is less obvious which error type we should credit to the translation model. However, we argue that SCORE errors are lower on the error hierarchy than SENSE errors because the latter could change to the former as models are augmented but the opposite would never happen unless translations are removed from the model. Utilizing such an error hierarchy is reasonable in our experimental settings, where our comparison systems are always supersets of the baselines; translation pairs and scores are never removed from baseline models. We acknowledge that such an approach may not be appropriate in all settings, however.

Next, consider the first word in the test sentence, *él*, which is aligned with *he* in the first reference and with *he is* in the second. The first output makes one SENSE error with respect to the first reference and two SENSE errors with respect to the second. In combining references, it is not clear whether the output should be marked as having made one or two sense errors. This example illustrates how alignment errors make the analysis difficult to do accurately.

Our approach to doing WADE analyses with multiple references slightly redefines the unit of analysis. Instead of using test-reference alignment links initially, we first consider each *word in the source language input*. For each word in a given test sentence, we identify the reference for which the set of WADE-annotated links out of the test word are 'best.' As mentioned, we use a hierarchy of error types to define those that are more egregious than others (i.e. SEEN errors are worse than SENSE errors, SENSE errors are worse than SCORE errors). Consider the first MT output in the example in Figure 6.8. The best set of annotated alignments from the first test set word, *él*, are taken from the first reference (a single SENSE error). The WADE-annotated alignments from the second word, *es*, are identical under both references (a single correct link). The third word, *muy*, is more accurately translated under the second reference (zero links in comparison with a SCORE error) and, finally, the alignment from the fourth word, *lindo*, is a SEEN error under both references.

This approach requires us to rigorously define how to compare two sets of annotated alignments. We employ the following strategy:

1. SEEN errors are assigned 4 points, SENSE errors are assigned 3 points, SCORE errors are assigned 2 points, and a correct alignment is assigned 0 points.

2. For each set of reference links, point values are averaged across all alignment links. If there are no word alignments to a given test set word, a null alignment point value of 1 is assigned.

For each input test word, we use this strategy to compare the alignment links between it and the set of aligned words in each of our reference translations and choose the lowest scoring set. The chosen set of 'best' alignment links for each input word are included in the cumulative WADE analysis, which, as in the original formulation of WADE, is measured across a single set of test-reference alignment links. In the case that we only have one reference translation, the 'best' set of alignment links for each test word is always taken from the sole reference translation, and the resulting analysis follows our original definition of WADE, given in Section 3.4.1.2.

One drawback of this approach is that it does not reward longer ngram matches. That is, for example, the first word in an input sentence could be best matched with the first reference, the second word with the second reference, the third word with the third reference, etc. This is a general shortcoming of WADE, and it is only made more clear in the multiple reference version. However, this is the cost of being able to generate detailed error-type annotations.

6.4.4.2 Analysis

Table 6.6 shows the results of our multi-reference WADE analysis over the output from several of the low resource models described in Section 6.4.3. In particular, we present WADE results over the models represented by the first and last point in each of the four learning curves and six source languages shown in Figure 6.5.

In general, the results of the WADE analyses are what we might expect. When we add translations for OOV words (+ Translations models), the number of SEEN errors decrease substantially.[11] The corresponding increases in SENSE errors indicate that many of the previously SEEN errors become SENSE errors, which makes sense; our models have now seen the previously OOV words, but they do not include all correct translations because our bilingual lexicon induction method is imperfect. In most cases the number of correct alignments also increases, which is in keeping with the observed increases in BLEU score.

It would be insightful to be able to measure the shifts between error types across different outputs. For example, it appears that we when add our induced translations, many previously SEEN errors become SENSE errors and some are corrected. It would be interesting to see the frequency of each shift. Although this is straightforward to measure in a single reference setting, it is more complicated with multiple references. For example, in the output of a first SMT model a given source language test word may have a best set of alignment links with a reference that includes one SEEN error and one SENSE error. Then, in the output of a second SMT model, the same source language test word may have a best set of alignment links with a different refers that includes one SCORE error and one correct link. Because the two sets of best alignment links consider different reference translations, we cannot know whether the original SEEN

[11] A few SEEN errors remain for those words which appear so infrequently in our comparable corpora that we are not able to induce translations for them.

Language	Model	SEEN	SENSE	SCORE	Correct	BLEU
Bengali	500 sentence pairs baseline	48.0	23.0	4.9	24.1	2.30
	+ Scores	48.0	23.0	4.9	24.1	2.13
	+ Translations	2.3	64.8	4.0	28.9	3.50
	+ Scores and Translations	2.3	64.4	3.8	29.4	3.17
	Full training data	20.6	17.5	11.7	50.1	12.07
	+ Scores	20.6	17.6	11.9	49.9	12.25
	+ Translations	1.3	35.6	11.6	51.5	12.74
	+ Scores and Translations	1.3	35.6	11.6	51.6	12.55
Tamil	500 sentence pairs baseline	71.0	9.8	0.6	18.6	1.69
	+ Scores	71.0	9.8	0.6	18.6	1.71
	+ Translations	2.3	74.2	0.6	23.0	2.53
	+ Scores and Translations	2.3	74.2	0.6	22.9	2.49
	Full training data	26.3	20.9	10.1	42.7	9.45
	+ Scores	26.3	20.9	10.3	42.5	9.77
	+ Translations	1.6	44.2	10.0	44.1	9.45
	+ Scores and Translations	1.6	44.3	10.1	43.9	9.98
Telugu	500 sentence pairs baseline	69.5	14.4	8.2	7.9	2.04
	+ Scores	69.5	14.4	8.3	7.9	2.04
	+ Translations	5.5	64.9	3.7	25.9	4.18
	+ Scores and Translations	5.5	64.9	3.8	25.9	4.24
	Full training data	28.8	18.2	14.0	39.0	11.72
	+ Scores	28.8	18.2	13.9	39.1	11.96
	+ Translations	3.0	42.9	14.1	40.0	12.20
	+ Scores and Translations	3.0	42.9	13.9	40.3	12.25
Urdu	500 sentence pairs baseline	41.9	18.1	4.2	35.7	1.83
	+ Scores	42.0	18.1	4.4	35.6	2.26
	+ Translations	1.5	45.6	4.3	48.6	5.44
	+ Scores and Translations	1.5	45.6	4.4	48.4	5.43
	Full training data	8.7	7.1	9.5	74.7	20.39
	+ Scores	8.7	7.1	9.8	74.4	20.97
	+ Translations	0.6	14.0	9.4	76.0	21.30
	+ Scores and Translations	0.6	14.1	9.6	75.7	21.78
Hindi	500 sentence pairs baseline	30.2	12.4	5.8	51.6	4.23
	+ Scores	30.5	12.5	6.3	50.7	4.46
	+ Translations	1.4	35.0	5.8	57.8	6.05
	+ Scores and Translations	1.4	35.2	6.1	57.3	6.09
	Full training data	11.9	6.3	8.3	73.5	15.01
	+ Scores	11.9	6.3	8.4	73.4	15.34
	+ Translations	0.7	16.4	8.3	74.5	15.59
	+ Scores and Translations	0.7	16.3	8.1	74.8	16.08
Malayalam	500 sentence pairs baseline	79.8	7.0	1.9	11.3	0.5
	+ Scores	79.6	6.9	1.2	12.2	0.42
	+ Translations	4.1	75.5	5.4	14.9	1.21
	+ Scores and Translations	4.1	75.6	5.7	14.6	0.99
	Full training data	2.2	37.8	12.2	47.8	13.55
	+ Scores	2.2	37.7	11.9	48.1	14.15
	+ Translations	1.5	38.6	12.7	47.2	13.65
	+ Scores and Translations	1.5	38.5	12.4	47.6	14.18

Table 6.6: WADE Analysis of Augmented SMT Models. The percent of alignment links in the 'best' set (computed separately for each input source language word) that are annotated with each error type is given. For comparison, the multi-reference BLEU score on the output from each model is also given.

error or the original SENSE error has been corrected. To make such an inference we would need to align the reference translations with one another. We leave further research in this vein for future work.

In the results we presented in Section 6.4.3, when we added new phrase table features estimated over our comparable corpora, we typically observed small but consistent increases in BLEU scores, and we attributed these gains to a decrease in SCORE errors. However the results in Table 6.6 do not show consistently decreasing numbers of SCORE errors when we add the new features. As mentioned, WADE does not estimate the quality of multi-word ngrams. However, when we manually compare the outputs from our feature-augmented models with the outputs from our baseline models, we see that they tend to include longer ngram matches with the reference translations. That is, our new features seem to correct reordering errors more often than they correct lexical selection errors. This is possible in cases where the new feature functions give high scores to phrase translations, and the decoder is able to translate longer complete phrases. We can see this effect in the component ngram precisions that contribute to overall BLEU scores. For example, the Malayalam baseline model trained on the full dataset achieves a BLEU score of 13.55 and when we add our comparable corpora-based features, the BLEU score increases to 14.15. However, SCORE errors only decrease in number slightly. When we take a closer look at the BLEU score, we see that the unigram precision only increases by 3% but the bigram precision increases by 5%.

6.4.4.3 WADE Analysis Conclusions

WADE was originally meant to be a tool for doing error analysis, not for evaluating and comparing machine translations. The distinction is slight but significant. As an error analysis technique, it provides a nice way to visualize system outputs and gives a general sense of what types of lexical selection errors are made. However, using the method to evaluate and compare only slightly different models and their outputs is perhaps not appropriate. When we compare very similar machine translation outputs, errors in the automatic word alignments that WADE is based upon become relatively more substantial. Additionally, making the fine-grained distinctions and comparisons required for doing multi-reference WADE is unsatisfying as an overall measure of translation quality. Both error analysis and automatic methods for measuring quality are important and active subfields of machine translation, and we leave further improvements to methods like WADE to future work.

6.5 End-to-End SMT with Zero or Small Parallel Texts Conclusion

In this chapter, we applied bilingual lexicon induction techniques that we presented in Chapter 4 and phrase table scoring techniques that we presented in Chapter 5 to the settings where we have zero or only a small amount of parallel text for training SMT models. We focused on translating truly low resource languages. Our experiments showed the following:

1. In settings where we have only a dictionary of word translations, compiling a model that takes advantage of transliterations, feature functions estimated over comparable corpora, and induced translations increases the readability of outputs substantially. Our qualitative experiments show that beginning with even a very small parallel corpus instead of a dictionary of word translations further improves translation quality substantially.

2. When we begin with a small training bitext, our approaches to improving coverage (induced translations) and accuracy (new phrase table feature functions) improve BLEU scores, and their combined impact is nearly additive.

3. In additional experiments on low resource languages where we begin with a small amount of parallel training data, we found the following:

   (a) Inducing translations for low frequency words in addition to OOV words can improve BLEU scores modestly.

   (b) Augmenting baseline models with top-k induced translations where k ranges from 3 to 25 can improve performance beyond augmenting models with only top-1 translations.

   (c) The impact of our coverage and accuracy improvements is greatest when very little training data ($< 10,000$ sentence pairs) is used.

   (d) The size of our comparable corpora impacts the accuracy of our new feature functions more than the quality of our induced translations.

4. We presented a multi-reference version of WADE. Although the resulting analyses are insightful, we caution that the methodology should be used carefully and not in all experimental conditions.

In Chapter 4, we presented methods for inducing word translations. In Chapter 7, we use those techniques to induce multi-word translations. Doing so involves considerable additional challenges.

Chapter 7

Phrase Translation Mining

In this chapter, we move from inducing *word* translations to inducing *phrase* translations. Figure 7.1 shows the percent of n-grams in our Hindi and Spanish SMT tuning sets that appear in varying amounts of training data. It is clear from the plots that when only limited amounts of parallel training data are available, the corresponding phrase tables will have limited coverage, particularly for multi-word source phrases. In this chapter, we present our approach to inducing phrase translations beyond those that can be identified with small amounts of bitext.

Like our methods for *scoring* phrase pairs, we use bilingual lexicon induction in order to learn phrase translations from comparable corpora. If the source and target language each contain, for example, $100,000$ unigram word types, the number of pairwise comparisons is about 10 billion, which is significant but computationally feasible. In theory, we *could* follow our method for inducing word translations directly in order to induce multiword phrase translations. That is, we could score all source language phrases paired with all target language phrases using signals derived from comparable corpora and use the phrase pairs extracted from the small bitext as supervision for combining the signals in a discriminative classifier. However, in contrast to unigrams, the difficulty in inducing a comprehensive set of *phrase* translations is that the number of phrases, on both the source and target sides, is immense, and such a brute force approach is not computationally feasible. In theory, the number of possible phrases of up to length m in the target language is V_t^m, where V_t is the size of the target language unigram vocabulary. That is, a target language phrase, in theory, is any sequence of m words sampled from V_t. Similarly, the number of possible phrases of up to length m in the source language is V_s^m, where V_s is the size of the source language unigram vocabulary. Using a brute-force induction method comparable to our method for inducing unigram translations would require making $V_t^m * V_s^m$ phrase comparisons.

Of course, we don't observe all possible ordered combinations of words in a language in naturally occurring text. For example, outside of this document, the phrase "round moses phrases" is unlikely to occur even though all three words in the phrase are reasonably common.[1] However, even if we limit the phrase comparisons to those that we observe in monolingual corpora, the number of pairwise comparisons is computationally infeasible. For example, there are about 83 and 113 million unique phrases up to length three in the English and Spanish Wikipedias, respectively. The total number of pairwise comparisons necessary to do an exhaustive search over all source and target language phrases is over 9 quadrillion. Thus, even if we limit the task to short, observed phrases, the number of required pairwise phrase comparisons is infeasible.

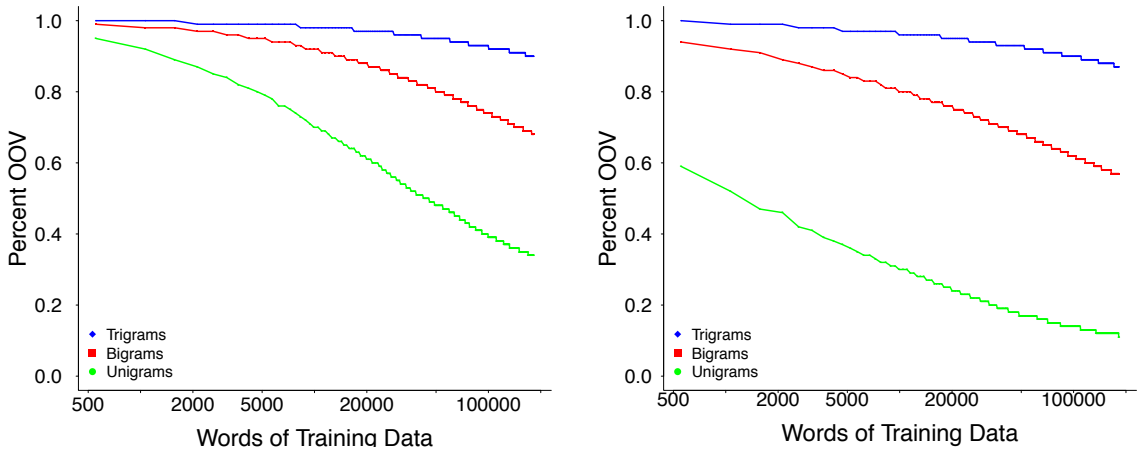
In addition to the computational challenges, doing an exhaustive pairwise search for phrase translations may not yield good results. Rapp and Sharoff (2014) showed, for example, that multiword phrases are too infrequent for a model based only on contextual information to make good predictions. This negative result was found even in the scenario where over half a billion words of source and target language Wikipedia data was used.

However, supplementing a low resource machine translation model with induced phrase translations is potentially quite useful. Multi-word translation units have been shown to improve the quality of SMT dramatically (Koehn *et al.*, 2003). Phrase translations allow translation models to memorize local context-dependent translations and reordering patterns.

There are several ways to reduce the computational complexity of this task as we have defined it, including the following:

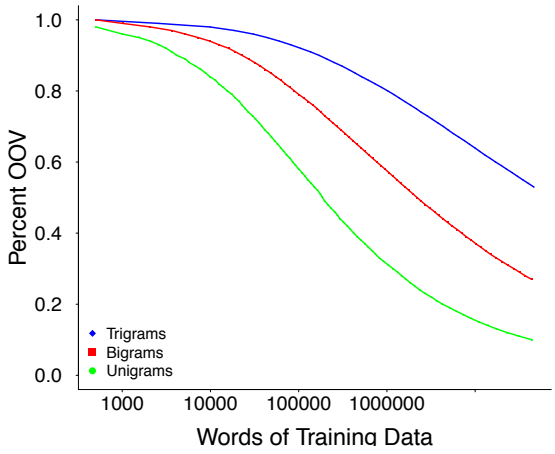
- Reduce the complexity of estimating monolingual signals

[1] In fact, a google search on the trigram yields zero results

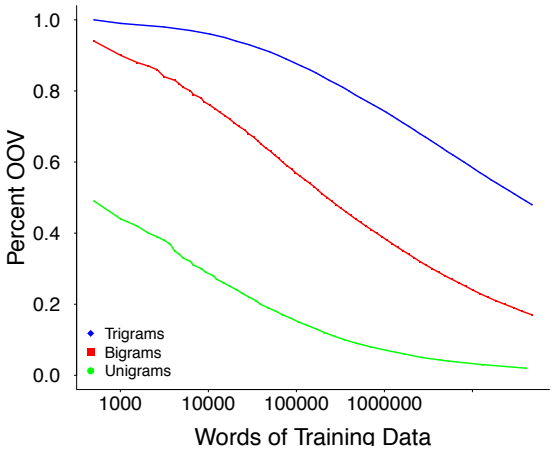


(a) **Hindi Type-Based**

(b) **Hindi Token-Based**



(c) **Spanish Type-Based**



(d) **Spanish Token-Based**

Figure 7.1: Hindi (a, b) and Spanish (c, d) type-based (a,c) and token-based (b, d) phrasal OOV rates for the SMT development set for each language over a varying number of training data words. The Spanish training data consists of about 50 million words of Europarl text, and the Hindi training data consists of about 180 thousand words of crowdsourced translations of Wikipedia documents.

  - Sample monolingual corpora
    - Uniformly, for building signatures of all phrases
    - Proportional to the frequency of a given phrase
  - Ignore dimensions that are unlikely to be informative (e.g. very low or high frequency words as contexts)
- Reduce the complexity of computing the similarity between pairs of monolingual signatures
  - Use randomized algorithms (e.g. locality sensitive hashing (Charikar, 2002; Goemans and Williamson, 1995; Indyk and Motwani, 1998)) to reduce the dimensionality of signatures
- Prune the set of phrase pair comparisons
  - Limit the set of source phrases for which to induce translations
  - Limit the set of target phrases which are eligible to be translations for any source phrase
  - For each source phrase, limit the set of target phrases to those that are more likely translations
  - Approximate the search over all target phrases by sorting signatures and comparing only the k-closest signatures in the sort, rather than all (e.g. using PLEB, or Point Location among Equal Balls techniques (Indyk and Motwani, 1998))

In this chapter, we focus on the major source of complexity: the number of pairwise phrase comparisons. Rather than compare *all* source language phrases with *all* target language phrases, we identify ways to propose a smaller set of hypothesis phrase translations for each source language phrase. In Section 7.1, we explore several ways to efficiently *filter* the search space of all target language phrases and in Section 7.2 we explore ways to *compose* target phrase translations from known unigram translations. Both sets of exploratory experiments motivate the novel algorithm that we propose in Section 7.2.2. We use features estimated over comparable corpora to rank candidate target phrase translations in Section 7.2.3, using a method similar to the one we developed in Chapter 4 for ranking unigram translations. Finally, in Section 7.3, we present experiments using induced phrase translations in end-to-end SMT. This chapter presents and extends the methods and experiments that we published in Irvine and Callison-Burch (2014a).

7.1 Option 1: Fast Phrase Pair Filtering

One way to reduce the the set of N^2 pairwise phrase comparisons is to do an initial, fast *pruning* of the set, proposing only a subset for full monolingual similarity scoring. We want the subsets of candidate phrase pairs to (i) include as many high quality phrase pairs as possible, and (ii) be small enough so that it is feasible to compute the monolingual similarity between all pairs in the set. That is, we would like both recall and precision on the set to be high. There is a tradeoff between including as many good pairs as possible (emphasizing a high recall) and limiting the size of the list (emphasizing a high precision), which we will discuss throughout our experiments.

The main intuition behind our approach is the following: there are some characteristics of given source phrases that may allow us to effectively and efficiently prune the space of all possible target phrase translations to a much smaller subset. For example, it may be true that source language bigrams are almost never translated as trigrams in the target language. Or it may be true that the number of stop words in a given source language phrase is usually the same as in its target language translation. Or, perhaps, given that we have access to a bilingual dictionary of *single word* translations, translations of source phrases are likely to contain at least one word translation that we already know about. Some *filters*, such as these, can be implemented efficiently as *inverted indices*. Figure 1 shows a naive approach to constructing a phrase table; all source language phrases are compared with all target language phrases. In contrast, the algorithm in Figure 2, only compares each source language phrase with target language phrases for which some given feature holds. In order to make filtering efficient, we would only want to allow features which can be pre-computed and stored in inverted indices. Doing so eliminates the need to iterate over the entire set of all possible target phrases for each source phrase.

We may also want to allow the use of multiple features in order to prune the space of target phrases. For example, Algorithm 3 sketches a pruning model that chooses the subset of target phrases that have feature number 1 or both features 2 and 3, or both. This model could, for example, correspond to the following: for each source phrase, choose target phrases that contain a translation of at least one word in the source phrase and/or that have the same length and the same number of stop words. In this case, the function *getFeat1* would return a set of all of the target

```
for src in srcPhrases do
    for trg in trgPhrases do
        compare(src,trg);
    end
end
```

Algorithm 1: Unpruned comparison of src and trg source and target phrases.

```
for src in srcPhrases do
    feat = getFeat(src);
    for trg in featToTrgP[feat] do
        compare(src,trg);
    end
end
```

Algorithm 2: Pruning source and target phrase comparisons using features implemented as inverted indices.

```
for src in srcPhrases do
    feat1 = getFeat1(src);
    feat2 = getFeat2(src);
    feat3 = getFeat3(src);
    for trg in union(featToTrgP1[feat1] , intersection(featToTrgP2[feat2], featToTrgP3[feat3]) ) do
        compare(src,trg);
    end
end
```

Algorithm 3: Pruning source and target phrase comparisons using features implemented as inverted indices.

language translations of all unigrams in the source phrase, and the function *featToTrgP1* would, given a set of target language words, return a set of all phrases that contain at least one of those words. Similarly, *getFeat2* would return the length of the input source phrase, and *featToTrgP2* would return a set of all target phrases with a given length. The inverted indices require a fixed amount of time to construct, which depends on the number of target language phrases and the complexity of the feature computation. Additionally, they have the potential to require large amounts of memory if a given target phrase is associated with many feature values. For example, the *featToTrgP* inverted index associates each target phrase with all translations of each of its component unigrams. In this work, we largely ignore memory requirements and aim to minimize time complexity. Computing the intersection of two sets, a and b, requires O(min(len(a),len(b))) time. Therefore, we hope to only use intersected features if at least one of them typically returns a relatively small set of target phrases.

In Appendix 15 we present a set of detailed experiments in which we automatically learn filters for very quickly pruning the space of hypothesis target language phrase translations. Fortunately, our experiments show that we can achieve both a high accuracy and a high filtering rate with a small number of filters, all of which can be implemented as inverted indices. Those experiments directly led to the algorithm that we propose in Section 7.2.

7.2 Option 2: Composing Phrase Translations

In Section 7.1, we cast the problem of reducing the number of pairwise comparisons necessary to induce a phrasal translation dictionary as one of *filtering*. In that top-down approach, we started with a very large set of candidate English phrase translations and learned ways to very quickly prune that space given some source phrase that we were interested in translating. In this section, we present a bottom-up approach where we begin with a source phrase and *compose* target language phrase translations using what we know about how the unigrams in the source phrase translate.

Our bottom-up, compositional method builds upon the notion that many phrase translations can be composed from the translations of its component words and subphrases. For example, Spanish *la bruja verde* translates into English as *the green witch*. Each Spanish word corresponds to exactly one English word. The phrase pair could be memorized and translated as a unit, or the English translation could be composed from the translations of each Spanish unigram.

Zens *et al.* (2012) found that only 2% of phrase pairs in German-English, Czech-English, Spanish-English, and

French-English phrase tables consist of multi-word source and target phrases and are non-compositional. That is, for these languages, the vast majority of phrase pairs in a given phrase table could be composed from smaller units. That work defines compositionality strictly: a phrase pair is defined as compositional if there exists a set of smaller phrase pairs that can be used together to produce the same phrase translation. For example, if a model contains the translation from Spanish *gobierno* to English *government*, from Spanish *de* to English *of*, and from Spanish *Francia* to English *France*, then the larger phrase translation from Spanish *gobierno de Francia* to English *government of France* could be composed. In contrast, the English translation *French government* could *not* be composed unless there were additional rules allowing, for example, *de Francia* to translate as *French*.

Our approach here takes advantage of the fact that many phrases can be translated compositionally. However, in order to achieve high recall in our set of hypothesis translations, we explore looser definitions of compositionality than is typical. In particular, we experiment with ignoring stop words in both the source and target phrases and using unigram prefix stem translations to allow for unseen suffixal variations.[2]

7.2.1 Motivating Experiments

We formulate several experiments to answer these research questions:

1. What percent of phrase translations could be correctly translated by *composition*, using known unigram translations (a bilingual dictionary)?
2. Would a low resource translation model benefit from composing its unigram translations into phrases?
3. Would a low resource translation model benefit further from also composing translations using a bilingual lexicon learned from monolingual texts?

We define the process of compositional translation as follows:

1. Given a source phrase, identify all known unigram translations of each source word.
2. Enumerate all *combinations* of per-word dictionary translations.[3]
3. Enumerate all *permutations* of all combinations.[4]

Given this definition of compositional translation, we ask our first question: For a given language pair, what percent of phrase translations could be correctly translated by *composition*, using known unigram translations (bilingual dictionary)? That is, if we only had access to unigram translations, what percent of phrases could we translate correctly through composition? To answer this question, we use the same sets of phrase translations extracted from manually aligned development set Spanish and Hindi sentence pairs that we used in Appendix 15.2 (500 sentence pairs each; see Appendix 12). There are about $6,000$ and $3,000$ bigram and trigram translations in the Spanish and Hindi datasets, respectively. Here, we ask how many of the phrase pairs could be generated by compositional translation using each of the following:

1. Unigram translations extracted from a seed bitext
2. Unigram translations extracted from a seed bitext, with optional stop word insertion and deletion
3. Unigram translations extracted from a seed bitext and induced from comparable corpora, with optional stop word insertion and deletion

For both dictionary types, we use five character prefix stems, and we use the top-5 induced unigram translations for each source language word type in our induced dictionary. We compute statistics using initial unigram translations extracted from $1,000$, $2,000$, $4,000$, and $8,000$ randomly sampled lines of the Spanish-English Europarl bitext and the Hindi-English parallel corpus as well as each complete corpus. For comparison, we also compute the percent of test set phrase pairs that are reachable given each word aligned training corpus. Some pairs that are unreachable using our composition technique and word alignment unigram dictionary may be reachable under a different phrase pair

[2] E.g., if *nuestros* translates as *our*, we also consider *our* a translation of *nuestras*, since they share the prefix *nuest-*.

[3] e.g. If the first source word in a source bigram has two dictionary translations, t_{11} and t_{12}, and the second has three, t_{21}, t_{22}, and t_{23}, then there are six combinations of unigram translations: $t_{11}t_{21}$; $t_{12}t_{21}$; $t_{11}t_{22}$; $t_{12}t_{22}$; $t_{11}t_{23}$; and $t_{12}t_{23}$

[4] e.g. The pair $t_{11}t_{21}$ is permuted so that the final set of composed phrases contains both $t_{11}t_{21}$ and $t_{21}t_{11}$

	Test Data		Reachable via Training Data	Reachable via Composition			Total
Bitext Size	Src Phrases	Total Translations	In Training Bitext	Aligned Training	+ Stop Ins/Del	+ Induced Trans	Reachable
Spanish							
1k	5623	6468	5.0%	8.0%	10.6%	18.6%	21.7%
2k			7.1%	9.5%	13.7%	20.6%	24.6%
4k			10.0%	11.0%	16.6%	22.7%	28.1%
8k			12.8%	12.9%	19.9%	25.1%	31.6%
Full			40.5%	17.7%	28.5%	30.8%	51.0%
Hindi							
1k	2423	2841	2.8%	7.2%	11.0%	17.7%	19.1%
2k			5.3%	9.4%	14.9%	19.9%	22.5%
4k			7.7%	12.4%	18.5%	23.2%	27.0%
8k			11.3%	13.9%	21.4%	25.0%	30.4%
Full			18.2%	16.4%	25.0%	27.8%	36.7%

Table 7.1: Percent of bigram and trigram translations that could be composed from unigram translations. Corpus size is given in number of sentence pairs and determines the phrase table from which initial unigram translations are extracted. The first column of results gives the percent of test phrase pairs that are reachable using the training text alone, without composing any additional translations. The next set of results show the percent of phrase pairs that are reachable via compositional translation. The first column shows results using only the unigram translations in the phrase table for composition. The second column also allows for stop words to be deleted from the source (e.g. Given Spanish 'la paz en,' the 'la' can be ignored and 'peace in' qualifies as a compositional translation) or inserted in the English (e.g. Spanish 'negociación,' English 'negotiation' can be generated compositionally and 'the' can be inserted, giving 'the negotiation.'). The third column also uses unigram translations induced using our supervised bilingual lexicon induction technique. The final column of results shows the total percent of test set phrase pairs that could be covered by both standard extraction from training data and via translation composition.

extraction heuristic. For example, the grow-diag-final heuristic that we use allows for null aligned words in extracted pairs. Table 7.1 shows the compositionality statistics. Note that we only evaluate over multiword source phrases up to length three (bigrams and trigrams).

The results in Table 7.1 show that by adding composed phrase translations, we may more than double the number of high quality phrase translations in our models. For very low resource conditions, the number of reachable phrase translations increases from less than 10% to over 20%. For example, the Hindi-English baseline model trained on $2,000$ sentence pairs contains only 5.3% of the test set phrase translation pairs. However, when we use the phrase table and induced unigram translations to augment the baseline phrase table with compositional phrase translation, that number increases to 22.5%. Recall that the translations used for evaluation come from a manually word aligned held-out development set, not from the training data used to build initial translation models. A 100% coverage would mean that *all* translations for *all* phrases up to length three (defined as those which are consistent with the word alignments) in the development set were contained in the phrase table.

Another interesting result in the motivating experiments presented in Table 7.1 is that the compositionality rates for Spanish and Hindi are very similar. One may expect that because Spanish and English are more closely related, more phrases would be translated compositionally. However, our results do not show that. It would be interesting to do a similar analysis on a pair of completely unrelated languages (e.g. a non-Indo-European language). Because we do not have manual word alignments for such a language pair, we leave this for future work.

The composition algorithm that allows for stop word insertion and deletion and which uses dictionaries derived from both the aligned parallel corpora and induced unigram translations (the last column in Table 7.1) is **exactly equivalent** to the second decision tree presented in Appendix 15.2. Originally, the algorithm was presented as a top-down decision tree; here it has been presented as a bottom-up compositional algorithm. In Section 7.2.2, we define the efficient version of the algorithm formally, which uses inverted indexes as proposed in Section 7.1. However, we emphasize that the algorithm can be thought of either as a top-down filtering or as a bottom-up composition. In fact, our filtering experiments directly informed the way that we defined our composition algorithm. We showed in Appendix 15.2 that filters based on unigram translations were definitively more informative than those based on phrase

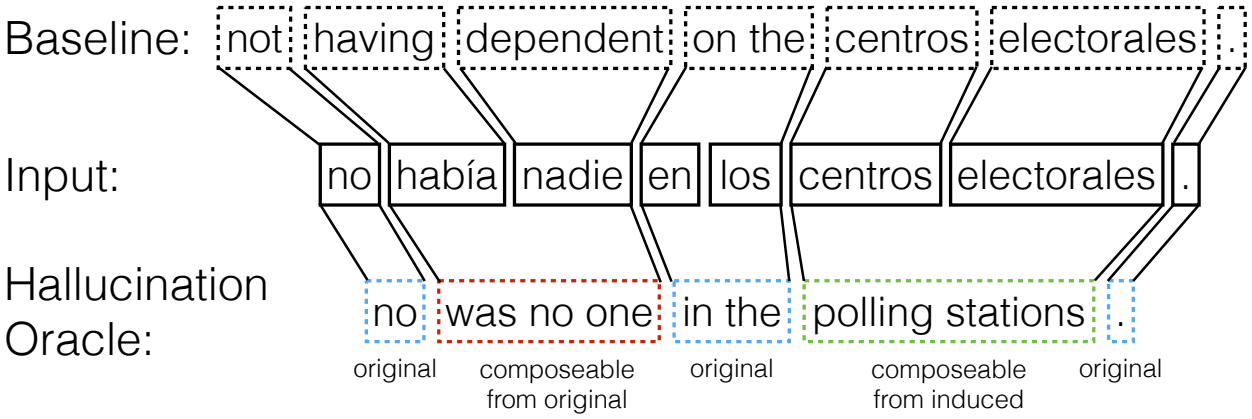


Figure 7.2: Example output from motivating experiment: a comparison of the baseline and full oracle translations of Spanish *no había nadie en los centros electorales*, which translates correctly as *there was nobody at the voting offices*. The full oracle is augmented with translations composed from the seed model as well as induced unigram translations. The phrase *was no one* is composeable from *había nadie* given the seed model. In contrast, the phrase *polling stations* is composeable from *centros electorales* using induced translations. For each translation, the phrase segmentations used by the decoder are highlighted.

length, monolingual frequencies and frequency differences, and we designed our algorithm for composing translations accordingly.

We now move to our second and third questions: Would an end-to-end low resource translation model benefit from composing its unigram translations into phrases? Would this be further improved by adding unigram translations that are learned from monolingual texts? Prior research has found that allowing an MT model to memorize longer translation units is crucial to achieving good performance. However, it's not clear how much of the observed performance gains are due to compositional phrase translations. There are two reasons for the gains from memorizing longer translation units; one involves compositional phrases and the other does not. First, memorizing longer translation units allows for the memorization of lexical choice and reordering patterns, including within compositional translations. For example, *ciudadanos cubanos* translates as *cuban citizens* and *centros electorales* translates as *polling stations*. In these cases, the Spanish phrases are translated compositionally, but the order of the two words is swapped in the English phrase. Additionally, the most frequent translation of Spanish *centros* is actually *centers*, however, in the context of *electorales*, *stations* is the preferred translation. Although these lexical choice and reordering patterns could be recovered by a high quality lexicalized or syntactic reordering model in combination with a language model, using longer translation units reduces the risk of making lexical choice and reordering errors.

Figure 7.2 illustrates how compositional phrase translations can result in improved translation quality. The baseline model is trained on a small amount of data, and it typically translates individual words instead of phrases. In a system augmented with compositional phrase translations, the translation *was no one* is composed from *habia nadie,* since *habia* translates as *was* in the baseline model, *nadie* translates as *one,* and *no* is a stop word. Similarly, the translation *polling stations* is composed from individual translations: *centros* translates as *stations*, and *electorales* translates as *polling*.

The second reason that memorizing longer translation units is beneficial is that non-compositional phrase translations may be learned. For example, content words may be left untranslated, as in *productos químicos* → *chemicals*, rather than *chemical products*. Similarly, *centros electorales* may be translated simply as *polls*. Or, more interestingly, phrases may be idiomatic, such as *maternidad remunerada* → *maternity leave*, where the unigram *remunerada* typically translates as *paid*, and the Spanish phrase *maternidad remunerada* as well as the English phrase *maternity leave* mean, idiomatically, 'the period during which a new mother is allowed paid time off.'

We answer our second and third questions by starting with low-resource Spanish-English and Hindi-English baselines trained from 2000 sentence pairs each (<60k words), and, for given sets of source phrases, augmenting each with (1) phrasal translations composed from its unigram translations, and (2) phrasal translations composed of a mix of the baseline unigram translations plus the monolingually-induced unigrams. We also add a set of monolingually-estimated feature functions.

Experiment	BLEU	
	Baseline Features	Monolingually Estimated Feats.
Spanish		
Low Resource Baseline	13.47	13.35
+ Oracle compositional trans. from training	14.90	15.18
+ Oracle compositional trans. from training and induced	15.47	15.94
Hindi		
Low Resource Baseline	8.49	8.26
+ Oracle compositional trans. from training	9.12	9.54
+ Oracle compositional trans. from training and induced	10.09	10.19

Table 7.2: Motivating Experiment: BLEU results using the baseline SMT model and composeable oracle translations with and without induced unigram translations.

For our motivating experiment we use oracle translations. Composed translations were only added to the phrase table if they were contained in the reference. This eliminated the huge number of noisy translations that are created by taking all of the permutations and combinations that arise when composing translations (we address this issue without using an oracle in Section 7.2.3).

In Section 7.3.3, we describe the set of source language phrases for which we compose translations. In the oracle experiments here, we augment a baseline model with translations for the same set of source language phrases. We use GIZA++ to word align our tuning and test sets. For both languages, we learn an alignment over our tuning and test sets and complete parallel training sets, and use a standard phrase pair extraction heuristic[5] to identify oracle phrase translations. We add oracle translations to each baseline model *without* bilingually estimated translation scores[6] because such scores will not be available for our automatically induced translations. Instead, as Section 7.2.3 details, we estimate 30 new phrase table features using comparable corpora. Thus, in our oracle experiments, we also measure performance using oracle phrase pairs scored with these features.

Table 7.2 shows the results of our oracle experiments. Augmenting the baselines with the subset of oracle translations which are *composed* given the unigram translations in the baseline models themselves (i.e. in the small training sets) yields a BLEU score improvement of about 1.4 points for Spanish and about 0.6 for Hindi. This finding itself is noteworthy, and we investigated the reason for it. A representative example of a compositional oracle translation that was added to the Spanish model is *para evitarlos*, which translates as *to prevent them*. In the training corpus, *para* translates far more frequently as *for* than *to*. Thus, it is useful for the translation model to know that, in the context of *evitarlos*, *para* should translate as *to* and not *for*. Additionally, *evitarlos* was observed only translating as the unigram *prevent*. The small model fails to align the adjoined clitic *los* with its translation *them*. However, our loose definition of compositionality allows the English stop word *them* to appear anywhere in the target translation.

In the first result, composeable translations do not include those that contain new, induced word translations. Using the baseline model *and* induced unigram translations to compose phrase translations results in a 2 and 1.6 BLEU point gain for Spanish and Hindi, respectively.

In Section 7.2.3, we describe the feature set that we use to rank hypothesized English translations for each Spanish and Hindi phrase. The second column of Table 7.2 shows the results of augmenting the baseline models with the same oracle phrase pairs as well as the new features estimated over *all* phrase pairs. Although the features do not improve the performance of the baseline models, this diverse set of scores improves performance dramatically when new, oracle phrase pairs are added. Adding all oracle translations and the new feature set results in a total gain of about 2.6 BLEU points for Spanish and about 1.9 for Hindi. These gains are the maximum that we could hope to achieve by augmenting models with our composed translations and new feature set.

To recap, we had asked the following motivating questions:

1. What percent of phrase translations could be correctly translated by *composition*, using known unigram translations (a bilingual dictionary)?

2. Would a low resource translation model benefit from composing its unigram translations into phrases?

[5] grow-diag-final

[6] We use an indicator feature for distinguishing new composed translations from bilingually extracted phrase pairs.

3. Would a low resource translation model benefit further from also composing translations using a bilingual lexicon learned from monolingual texts?

In answer to the first question, we found that, for both Spanish and Hindi, between about 20% and 30% of test set phrases could be translated by composition using known unigram translations. The results vary depending on the initial bilingual dictionary and the composition algorithm used. For the second question, we found that low resource Spanish and Hindi models improved by 1.4 and 0.6 BLEU points, respectively, when we augmented baselines with phrase translations composed from unigrams contained in the baseline models themselves. When we also added phrase translations composed from induced unigram translations (third question), we observed BLEU score increases of 2 and 1.6 for Spanish and Hindi, respectively.

7.2.2 Phrase Composition Algorithm

We have motivated our approach to *loosely composing* phrasal translations from the perspective of pruning the search space of all target phrases as well as from the perspective of composing translations from an existing unigram dictionary. Henceforth we will refer to our general approach as one of *composition*. A formal definition of our algorithm for composing phrase translations is shown in Algorithm 4, and an example translation is composed in Figure 7.3. Given a source language phrase, our approach considers all *combinations* and all *permutations* of all unigram translations for each source phrase content word. We ignore stop words in the input source phrase and allow any number of stop words anywhere in the output target phrase. As discussed in Section 7.1, we pre-compute an inverted index that maps sorted target language content words to sets of phrases containing those words in any order along with, optionally, any number of stop words. Our algorithm for composing candidate phrase translations is given in Algorithm 4, and an example translation is composed in Figure 7.3. Although in our experiments we compose translations for source phrases up to length three, the algorithm is generally applicable to any set of source phrases of interest.

Algorithm 4 yields a set of target language translations for any source language phrase for which all content unigrams have at least one known translation. As in our filtering experiments in 15.2, in an initial pruning step, we add a monolingual frequency cutoff to the composition algorithm and only add target phrases that have a frequency of at least θ_{Freq_T} to the inverted index. Doing so eliminates improbable target language constructions early on, for example *house handsome her* or *cute a house*. The size of the set of output translations depends on θ_{Freq_T}, the input phrase, and the unigram dictionaries.

7.2.3 Pruning Phrase Pairs Using Scores Derived from Comparable Corpora

We further prune the large, noisy set of hypothesized phrase translations before augmenting the seed translation model. To do so, we use a supervised setup very similar to that used for inducing unigram translations in Chapter 4; we estimate a variety of signals that indicate translation equivalence, including temporal, topical, contextual, and string similarity. As we showed in Chapter 5, such signals are effective for identifying phrase translations as well as unigram translations. In our experiments here, we add ngram length, alignment, and unigram translation features that are specific to our framework of phrase translation composition:

- Web crawl phrasal context similarity score
- Web crawl lexical context similarity score, averaged over aligned unigrams
- Web crawl phrasal temporal similarity score
- Web crawl lexical temporal similarity score, averaged over aligned unigrams
- Wikipedia phrasal context similarity score
- Wikipedia lexical context similarity score, averaged over aligned unigrams
- Wikipedia phrasal topic similarity score
- Wikipedia lexical topic similarity score, averaged over aligned unigrams
- Normalized edit distance, averaged over aligned unigrams
- Absolute value of difference between the logs of the source and target phrase Wikipedia monolingual frequencies
- Log target phrase Wikipedia monolingual frequency
- Log source phrase Wikipedia monolingual frequency
- Indicator: source phrase is longer
- Indicator: target phrase is longer
- Indicator: source and target phrases same length

Input: A set of source language phrases of interest, **S**, each consisting of a sequence of words $s_1^m, s_2^m, ...s_i^m$; A list of all target language phrases, $targetPhrases$; Source and target stop word lists, $Stop_{src}$ and $Stop_{trg}$; Set of unigram translations, $\mathbf{t}_{s_i^m}$, for all source language words $s_i^m \notin Stop_{src}$; monolingual target language phrase frequencies, $Freq_T$; Monolingual frequency threshold θ_{Freq_T}
Output: $\forall\ S^m \in \mathbf{S}$, a set of candidate phrase translations, $T_1^m, T_2^m, ...T_k^m$

Construct TargetInvertedIndex:
for $T \in targetPhrases$ **do**
  if $Freq_T(T) \geqslant \theta_{Freq_T}$ **then**
    T' ←words $t_j \in T$ if $t_j \notin Stop_{trg}$
    T'_{sorted} ← sorted(T')
    append T to TargetInvertedIndex[T'_{sorted}]
  end
end

for $S^m \in$ **S do**
  S' ←words $s_i^m \in S^m$ if $s_i^m \notin Stop_{src}$
  Combs$_{S'}$ ← $\mathbf{t}_{s'_1} \times \mathbf{t}_{s'_2} \times ... \times \mathbf{t}_{s'_k}$
  $T \leftarrow [\,]$
  for $c_{s'} \in$ *Combs*$_{S'}$ **do**
    $c_{s'_{sorted}}$ ← sorted($c_{s'}$)
    $T \leftarrow T$+TargetInvertedIndex($c_{s'_{sorted}}$)
  end
  $T^m = T$
end

Algorithm 4: Computing a set of candidate compositional phrase translations for each source phrase in the set S. An inverted index of target phrases is constructed that maps sorted lists of content words to phrases that contain those content words, as well as optionally any stop words, and have a frequency of at least θ_{Freq_T}. Then, for a given source phrase S^m, stop words are removed from the phrase. Next, the cartesian product of all unigram translations is computed. Each element in the product is sorted and any corresponding phrases in the inverted index are added to the output.

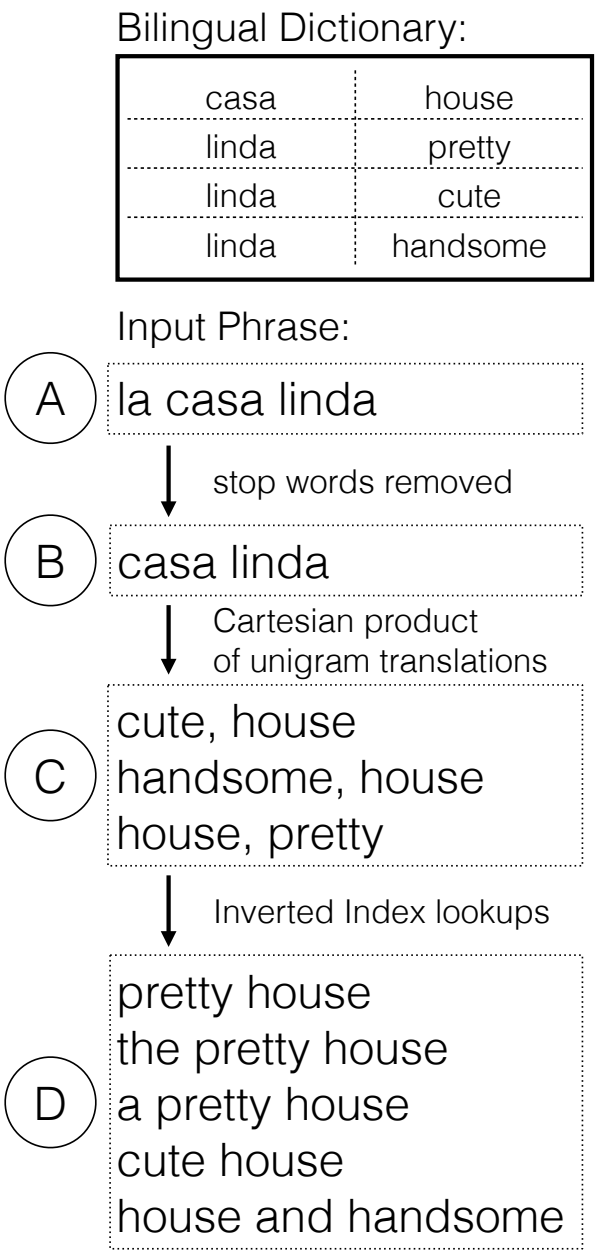


Figure 7.3: Example of loosely composed translations for the Spanish input in A, *la casa linda*. In B, we remove the stop word *la*. Then, in C, we enumerate the cartesian product of all unigram translations in the bilingual dictionary and sort the words within each alphabetically. Finally, we look up each list of words in C in the inverted index, and corresponding target phrases are enumerated in D. The inverted index contains all phrasal combinations and permutations of the word lists in C which also appear monolingually with some frequency and with, optionally, any number of stop words.

- Number of source content words higher than target
- Number of target content words higher than source
- Number of source and target content words same
- Number of source stop words higher than target
- Number of target stop words higher than source
- Number of source and target stop words same

- Percent of source words aligned to at least one target word
- Percent of target words aligned to at least one source word
- Percent of source content words aligned to at least one target word
- Percent of target content words aligned to at least one source word
- Percent of aligned word pairs aligned in bilingual training data
- Percent of aligned word pairs in induced dictionary
- Percent of aligned word pairs in stemmed induced dictionary

As we did for inducing unigram translations in Chapter 4, we learn a log-linear model for combining the features into a single score for predicting the quality of a given phrase pair. We extract training data from our baseline translation models trained on our small parallel corpora. We rank hypothesis translations for each source phrase using classification scores and keep the top-k. We found that using a score threshold sometimes improves precision. However, as experiments below show, the recall of the set of phrase pairs is more important, and we did not observe improvements in translation quality when we used a threshold.

7.3 End-to-end SMT with Induced Phrase Translations

7.3.1 Experimental Setup

In our SMT experiments here, we assume our typical data setting of having access to only a small parallel corpus. For our Spanish experiments, we randomly sample $2,000$ sentence pairs (about $57,000$ Spanish words) from the Spanish-English Europarl v5 parallel corpus (Koehn, 2005). For our Hindi experiments, we use the parallel corpora released by Post *et al.* (2012). Again, we randomly sample $2,000$ sentence pairs from the training corpus (about $39,000$ Hindi words). Additionally, we use approximately $2,500$ and $1,000$ single-reference parallel sentences each for tuning and testing our Spanish and Hindi models, respectively. Spanish tuning and test sets are newswire articles taken from the 2010 WMT shared task (Callison-Burch *et al.*, 2010).[7] We use the Hindi development and testing splits released by Post *et al.* (2012).

7.3.2 Unigram Translations

Of the $16,269$ unique unigrams in the source side of our Spanish MT tuning and test sets, 73% are OOV with respect to our training corpus. 21% of unigram tokens are OOV. For Hindi, 61% of the $8,137$ unique unigrams in the tuning and test sets are OOV with respect to our training corpus, and 18% of unigram tokens are OOV. However, because automatic word alignments estimated over the small parallel training corpora are noisy, we use bilingual lexicon induction to induce translations for *all* unigrams. We use our Wikipedia and online news web crawls comparable corpora to estimate similarity scores. Together, the two datasets contain about 900 million words of Spanish data and about 50 million words of Hindi data. For both languages, we limit the set of hypothesis target unigram translations to those that appear at least 10 times in our comparable corpora.

We use $3,000$ high probability word translation pairs extracted from each parallel corpus as positive supervision and $9,000$ random word pairs as negative supervision. We use Vowpal Wabbit[8] for learning. The top-5 induced translations for each source language word are used as both a baseline set of new translations (Section 7.3.5.3) and for composing phrase translations.

7.3.3 Composing and Pruning Phrase Translations

There are about 183 and 66 thousand unique bigrams and trigrams in the Spanish and Hindi tuning and test sets, respectively. However, many of these phrases do not demand new hypothesis translations. We do not translate those which contain numbers or punctuation. Additionally, for Spanish, we exclude names, which are typically translated identically between Spanish and English.[9] We exclude phrases which are sequences of stop words only. Additionally, we exclude phrases that appear more than 100 times in the small training corpus because our seed phrase table likely

[7] *news-test2008* plus *news-syscomb2009* for tuning and *newstest2009* for testing.

[8] `http://hunch.net/~vw/`, version 6.1.4. with standard learning parameters

[9] Our names list comes from page titles of Spanish Wikipedia pages about people. We iterate through years, beginning with 1AD, and extract names from Wikipedia 'born in' category pages, e.g. '2013 births,' or 'Nacidos en 2013.'

already contains high quality translations for them. Finally, we exclude phrases that appear fewer than 20 times in our comparable corpora as our comparable-corpora based features are unreliable when estimated over so few tokens. We hypothesize translations for the approximately $15,000$ and $6,000$ Spanish and Hindi phrases, respectively, which meet these criteria. Our approach for inducing translations straightforwardly generalizes to any set of source phrases.

In defining loosely compositional phrase translations, we use both the induced unigram dictionary and the dictionary extracted from the word aligned parallel corpus. We expand these dictionaries further by mapping unigrams to their five-character word prefixes. We use our Wikipedia comparable corpora to construct stop word lists, containing the most frequent 300 words in each language, and indexes of monolingual phrase frequencies. Recall that there are about 83 and 38 million unique phrases up to length three in the English sides of the Spanish and Hindi comparable corpora. However, we use a target monolingual frequency filter and ignore those target phrases that appear fewer than twenty times in the Spanish set and fewer than ten times in the Hindi set, reducing the size of each to 1.3 and 1.1 million English phrases, respectively. On average, our Spanish model yields $7,986$ English translations for each Spanish bigram, and $9,231$ for each trigram, or less than 0.1% of all possible candidate English phrases. Our Hindi model yields even fewer candidate English phrases, 826 for each bigram and $1,113$ for each trigram, on average.

We use the features enumerated in Section 7.2.3 to rerank candidate target translations for each source phrase. We extract supervision from the seed translation models by first identifying phrase pairs with multi-word source strings, that appear at least three times in the training corpus, and that are composeable using unigram translations in the models and induced dictionaries. Then, for each language pair, we use the $3,000$ that have the highest $p(f|e)$ scores as positive supervision. We randomly sample $9,000$ compositional phrase pairs from those not in each phrase table as negative supervision. As with inducing unigram translations, we use Vowpal Wabbit for learning a log linear model to score any phrase pair and rerank candidate target translations by their classification scores.

7.3.4 MT Experimental Setup

We use GIZA++ to word align each training corpus. We use the Moses SMT framework (Koehn *et al.*, 2007) and the standard phrase-based MT feature set, including phrase and lexical translation probabilities and a lexicalized reordering model. When we augment our models with new translations, we use the average reordering scores over all bilingually estimated phrase pairs. We tune all models using batch MIRA (Cherry and Foster, 2012). We average results over three tuning runs and use approximate randomization to measure statistical significance (Clark *et al.*, 2011).

For Spanish, we use a 5-gram language model trained on the English side of the complete Europarl corpus and for Hindi a 5-gram language model trained on the English side of the complete training corpus released by Post *et al.* (2012). We train our language models using SRILM with Kneser-Ney smoothing. Our baseline models use a phrase limit of three, and we augment them with translations of phrases up to length three in our experiments.

7.3.5 Results

7.3.5.1 Unigram Translations

Table 7.3 shows examples of top ranked translations for several Spanish words. Although performance is generally quite good, we do observe some instances of false cognates, for example the top ranked translation for *aburridos*, which translates correctly as *bored*, is *burritos*. Using automatic word alignments as a reference, we find that 44% of Spanish tuning set unigrams have a correct translation in their top-10 ranked lists and 62% in the top-100. For Hindi, 31% of tuning set unigrams have a correct translation in their top-10 ranked lists and 43% in the top-100. These results are slightly lower than those reported in Chapter 4, where we observed a 43% top-10 and 60% top-100 accuracy. The difference is likely due to the lower frequency and burstiness of the average word *type* in the tuning data.

7.3.5.2 Composed Phrase Pairs

Before moving to end-to-end SMT experiments, we evaluate the goodness of the composed and pruned phrase pairs themselves. In order to do so, we use the same set of oracle phrase translations described in Section 7.2.1.[10]

Table 7.4 shows the top three English translations for several Spanish phrases along with their model scores. Common, loose translations of some phrases are scored higher than less common but literal translations. For example, *very obvious* scores higher than *very evident* as a translation of Spanish *muy evidentes*. Similarly, *dutch minister* is scored higher than *netherlands minister* as a translation for *ministro neerlandès*.

[10] We do not use our manually aligned sets here because we only have annotations for a subset of the development and test sets.

Spanish	abdominal	abejorro	abril	aburridos	accionista	aceite	actriz
Top 5 English Trans.	**abdominal**	**bumblebees**	**april**	burritos	actionists	adulterated	**actress**
	abdomen	bombus	march	**boredom**	actionist	iooc	actor
	bowel	xylocopa	june	agatean	telmex	olive	award
	appendicitis	ilyitch	july	burrito	**shareholder**	milliliters	american
	acute	bumble	december	poof	antagonists	canola	singer

Table 7.3: Top five induced translations for several source words. Correct translations are bolded. *aceite* translates as *oil*.

Spanish	English	Score
ambos partidos	**two parties**	5.72
	both parties	5.31
	and parties	3.16
televisión estatal	state television and	8.12
	state television	7.13
	television interview	4.24
muy poderoso	**very powerful**	3.50
	powers were	1.77
	were powered	1.61
había apoyado	were supported	4.80
	were members	4.52
	had supported	4.39
ministro neerlandès	finnish minister	4.76
	finnish ministry	2.77
	dutch minister	1.31
unas cuantas semanas	over a week	4.30
	a few weeks	3.72
	few weeks	3.22
muy evidentes	**very obvious**	1.88
	very evident	1.87
	obviously very	1.84
prácticamente imposible	**almost impossible**	5.43
	almost all	3.52
	almost impossible to	2.22

Table 7.4: Top three compositional translations for several source phrases and their scores generated by our supervised discriminative model. Correct translations are bolded.

We use model scores to rerank candidate translations for each source phrase and keep the top-*k* translations. Figure 7.4 shows the precision and type-based recall (the percent of source phrases for which at least one correct translation is generated) as we vary *k* for each language pair. At $k = 1$, precision and recall are about 27% for Spanish and about 25% for Hindi.[11] At $k = 100$, recall increases to 57% for Spanish and precision drops to 2%. For Hindi, recall increases to 40% and precision drops to 1%.

Moving from $k = 1$ to $k = 100$, precision drops at about the same rate for the two source languages. However, recall increases less for Hindi than for Spanish. We attribute this to two things. First, our oracle experiments in Section 7.2.1 showed that there is less to gain in composing phrase translations for Hindi than for Spanish. Second, the accuracy of our induced unigram translations is lower for Hindi than it is for Spanish. Without accurate unigram translations, we are unable to compose high quality phrase translations.

Because we composed translations for source phrases that appear in the training data up to 100 times, our baseline model includes some of the oracle phrase translations. Not surprisingly, the bilingually extracted phrase pairs have

[11] Since we are computing type-based recall, and at k=1, we produce exactly one translation for each source phrase, precision and recall are the same.

Experiment	BLEU	
	Spanish	Hindi
Baseline	13.47	**8.49**
+ Mono. Scores	13.35	8.26
+ Mono. Scores & OOV Trans	**14.01**	8.31
+ Phrase Trans, k=1	13.90	8.16
+ Phrase Trans, k=2	14.07	8.86*
+ Phrase Trans, k=5	14.30*	8.89*
+ Phrase Trans, k=25	14.50*	9.00*
+ Phrase Trans, k=200	**14.57***	**9.04***

Table 7.5: Experimental results. First, the baseline models are augmented with monolingual phrase table features and then also with the top-5 induced translations for all OOV unigrams. Then, we append the top-k composed phrase translations to the third baseline models. BLEU scores are averaged over three tuning runs. We measure the statistical significance of each +Phrase Trans model in comparison with the highest performing (bolded) baseline for each language; * indicates statistical significance with $p < 0.01$ over the baseline performance.

high precision (81% and 40% for Spanish and Hindi, respectively) and low recall (6% and 15% for Spanish and Hindi, respectively).

7.3.5.3 End-to-End Translation

Table 7.5 shows end-to-end translation BLEU score results. Our first baseline SMT models are trained using only $2,000$ parallel sentences and no new translation model features. Our Spanish baseline achieves a BLEU score of 13.47 and our Hindi baseline a BLEU score of 8.49. When we add the 30 new feature functions estimated over comparable monolingual corpora, performance is slightly lower, 13.35 for Spanish and 8.26 for Hindi. Our third baselines augment the second with unigram translations for all OOV tuning and test set source words. We append the top-5 translations for each,[12] score both the original and the new phrase pairs with the new feature set, and retune. With these additional unigram translations, performance increases to 14.01 for Spanish and 8.31 for Hindi.

We append the top-k composed translations for the source phrases described in Section 7.3.1 to the third baseline models. Both original and new phrase pairs are scored using the new feature set. BLEU score results are shown at different values of k along the precision-recall plots for each language pair in Figure 7.4 as well as in Table 7.5. We would expect that higher precision and higher recall would benefit end-to-end SMT. As usual, a tradeoff exists between precision and recall, however, in this case, improvements in recall outweigh the risk of a lower precision. As k increases, precision decreases but both recall and BLEU scores increase. For both Spanish and Hindi, BLEU score gains start to taper off at k values over 25.

In additional experiments, we found that **without** the new features the same sets of composed phrase pairs hurt performance slightly in comparison with the baseline augmented with unigram translations, and results don't change as we vary k.[13] Thus, the translation models are able to effectively use the higher recall sets of new phrase pairs because we also augmented the models with 30 new feature functions, which help them distinguish good from bad translations.

7.3.6 Discussion

Our results showed that including a high recall set of composed translations in our augmented phrase table successfully improved the quality of our machine translations. The algorithm that we proposed for hypothesizing translations is flexible, and in future work we plan to modify it slightly to output even more candidate translations. For example, we could retrieve target phrases which contain at least one source word translation instead of all. Alternatively, we could identify candidates using entirely different information, for example the monolingual frequency of a source and target word, instead of unigram translations. This type of inverted index may improve recall in the set of hypothesis phrase translations at the cost of generating a much bigger set for reranking.

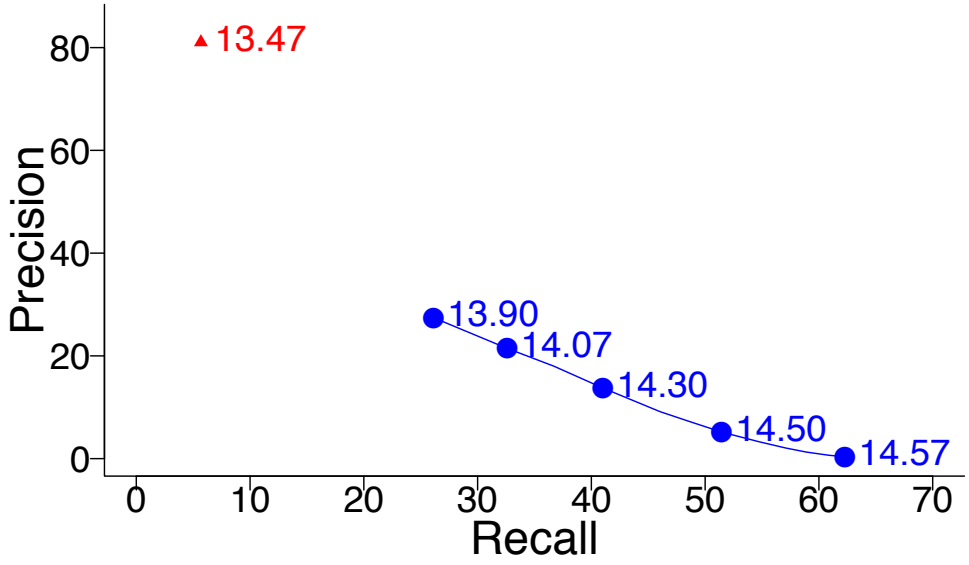
[12] The same set used for composing phrase translations.

[13] For all values of k between 1 and 100, without the new features, BLEU scores are about 13.70 for Spanish

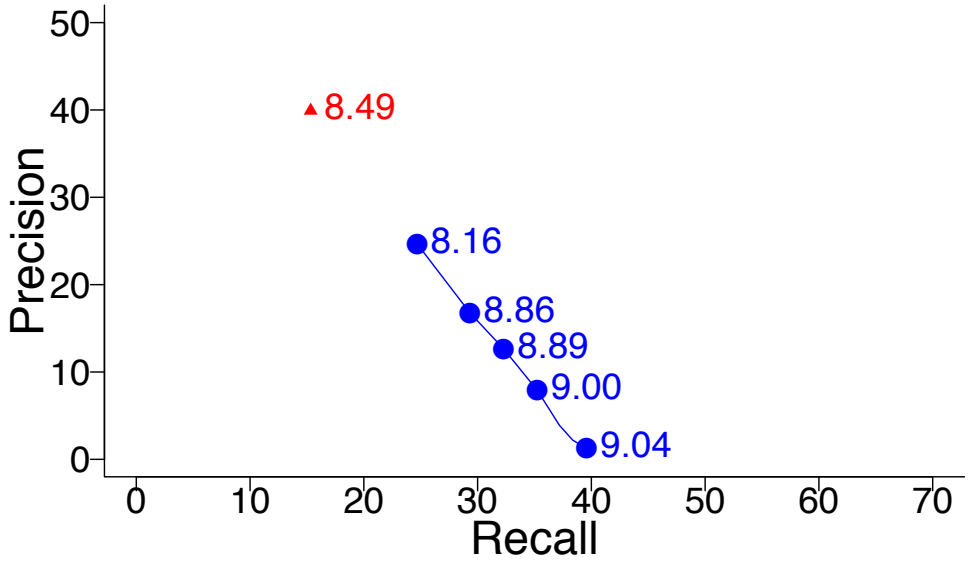
Our new phrase table features were informative in distinguishing correct from incorrect phrase translations, and they allowed us to make use of noisy but high recall supplemental phrase pairs. This is a critical result for research on identifying phrase translations from non-parallel text. We also believe that using fairly strong target (English) language models contributed to our models' ability to discriminate between good and bad composed phrase pairs. We leave research on the influence of the language model in our setting to future work.

In this work, we experimented with two language pairs, Spanish-English and Hindi-English. While Spanish and English are very closely related, Hindi and English are less related. Our oracle experiments showed potential for composing phrase translations for both language pairs, and, indeed, in our experiments using composed phrase translations we saw significant translation quality gains for both. We expect that improving the quality of induced unigram translations will yield even more performance gains.

The vast majority of prior work on low resource MT has focused on Spanish-English (Dou and Knight, 2012, 2013; Haghighi *et al.*, 2008; Klementiev *et al.*, 2012; Ravi, 2013; Ravi and Knight, 2011). Although such experiments serve as important proofs of concept, we believe it is important to also experiment with a more truly low resource language pair. The success of our approach that we have seen for Spanish and Hindi suggests that it is worth pursuing such directions for other even less related and resourced language pairs. In addition to language pair, text genre and the degree of looseness or literalness of given parallel corpora may also affect the amount of phrase translation compositionality.



(a) **Spanish**



(b) **Hindi**

Figure 7.4: Precision and recall curve with BLEU scores for the top-k scored composed translations. k varies from 1 to 200. Baseline model performance is shown with a red triangle..

Chapter 8

From Low Resource MT to Domain Adapted MT

In this chapter, we apply several of our techniques for improving the performance of machine translation for low resource language pairs to low resource text *domains*. We present and extend the approaches and results that we published in Irvine and Callison-Burch (2014b). Domain adaptation in machine translation is known to be a challenging research problem that has substantial real-world application. In this setting, we have access to training data in some old-domain of text but very little or no training data in the domain of the text that we wish to translate. For example, we may have a large corpus of parallel newswire training data but no training data in the medical domain, resulting in low quality translations at test time due to the mismatch.

In Section 3.4 and, originally, in Irvine *et al.* (2013a), we introduced a taxonomy for classifying machine translation errors related to lexical choice as well as two techniques, WADE and TETRA, for measuring the impact of each error type. Recall that our 'S4' taxonomy includes SEEN, SENSE, SCORE, and SEARCH errors. In Section 8.2, we present WADE and TETRA error analyses in a domain shift setting. We conclude that SEEN and SENSE errors are the most frequent but that there is also room for improving errors due to inaccurate translation model scores.

In this chapter, we use comparable corpora to reduce the number of SEEN and SCORE errors in a domain adaptation setting. In Carpuat *et al.* (2013), we defined the *SenseSpotting* task, where the goal is to identify words in a new-domain monolingual text that appeared in the old-domain text but which have a new, previously unseen sense. Although some sense shifts do not demand new translations, many do. If we could reliably identify words with new senses in new-domain text, then it may be possible to use our bilingual lexicon induction techniques to learn new translations for them and reduce SMT SENSE errors. We leave this for future work.

We assume the setting where we have an old-domain parallel training corpus but no new domain training corpus.[1] We do, however, have access to a mixed-domain comparable corpus. We identify new-domain text within our comparable corpus and use that data to (1) estimate new translation features on the translation models extracted from old-domain training data, and (2) induce translations for unseen words. Specifically, we focus on the French-English language pair because carefully curated datasets exist in several domains for tuning and evaluation. Following our prior work, we use the Canadian Hansard parliamentary proceedings as our old-domain and adapt models to both the medical and the science domains. At over 8 million sentence pairs, the Canadian Hansard dataset is one of the largest publicly available parallel corpora and provides a very strong baseline. In Irvine *et al.* (2013a), we showed that, unlike the newswire domain, the medical and science domains are very different from the parliamentary proceedings domain.

We give details about each dataset in Section 8.1. In Section 8.2, we present an error analysis of what goes wrong when we shift domains. Finally, in Sections 8.4 and 8.5, we describe novel methods for reducing, first, the number of SCORE errors and, then, SEEN errors in a domain adaptation setting.

8.1 Domain Adaptation Data

We assume that the following data is available in our translation setting:

[1] Some prior was has referred to old-domain and new-domain corpora as out-of-domain and in-domain, respectively.

Corpus	Source Words	Target Words
Training		
Canadian Hansard	161.7 m	144.5 m
Tune-1 / Tune-2 / Test		
Medical	53k / 43k / 35k	46k / 38k / 30k
Science	92k / 120k / 120k	75k / 101k / 101k
Language Modeling and Comparable Corpus Selection		
Medical	-	5.9 m
Science	-	3.6 m

Table 8.1: Summary of the size of each corpus of text used in this chapter in terms of the number of source and target word tokens.

- Large old-domain parallel corpus for training
- Small new-domain parallel corpora for tuning and testing
- Large new-domain English monolingual corpus for language modeling and identifying new-domain-like comparable corpora
- Large mixed-domain comparable corpus, which includes some text from the new-domain

These data conditions are typical for many real-world uses of machine translation. A summary of the size of each corpus is given in Table 8.1, and brief descriptions of the parallel datasets are below. In all cases, we use publicly available training, tuning, and test sets.[2]

Hansard: Canadian parliamentary proceedings, consists of manual transcriptions and translations of meetings of Canada's House of Commons and its committees from 2001 to 2009. Discussions cover a wide variety of topics, and speaking styles range from prepared speeches by a single speaker to more interactive discussions. It is significantly larger than Europarl, the common source of old domain data (Foster and Kuhn (2007), Koehn and Schroeder (2007), and Haddow and Koehn (2012), among others).

Medical: Medical documents from the European Medicines Agency (EMEA), made available with the OPUS corpora collection (Tiedemann, 2009). This corpus primarily consists of drug usage guidelines, which use boilerplate sentences that are often repeated across documents.

Science: Parallel abstracts from scientific publications in many disciplines including physics, biology, and computer science. The data was collected from two distinct sources: (1) Canadian Science Publishing made available translated abstracts from their journals which span many research disciplines; (2) parallel abstracts from PhD theses in Physics and Computer Science collected from the HAL public repository (Lambert *et al.*, 2012).

Parallel corpora are not available for the vast majority of text domain and language pair combinations in quantities large enough to train domain-specific models. In this chapter, we do not use any new-domain parallel data for training; we assume that we only have enough for tuning and testing. However, we do use the English side of the new-domain training corpora for language modeling and for identifying new-domain-like comparable corpora from our mixed domain comparable corpus. We address how we select new-domain comparable corpora in Section 8.3.

8.2 WADE and TETRA Analyses

Before applying our techniques for using comparable corpora to improve machine translation performance, we analyze what types of errors are made when we use a translation model trained on old domain data to translation text in some new domain. We extract a phrase-based model with a phrase limit of seven from the full Hansard parallel dataset. We use the standard phrase-based SMT feature set, including forward and backward phrase and lexical translation probabilities as well as a standard lexicalized reordering model. For each domain, we use two 5-gram language models:

[2] http://www.umiacs.umd.edu/~hal/damt/

Domain	% Correct			% Seen Errors			% Sense Errors			% Score Errors		
	OLD	MIXED	%Δ	OLD	MIXED	%Δ	OLD	MIXED	%Δ	OLD	MIXED	%Δ
Medical	55.97	62.60	+12%	9.28	4.01	-56%	16.16	13.76	-15%	18.59	19.63	+ 6%
Science	56.20	61.50	+9%	10.22	5.63	-45%	13.58	13.11	- 3%	20.00	19.77	- 1%

Table 8.2: WADE: Percent correct, percent seen errors, percent sense errors, and percent score errors. The changes (%Δ) from OLD to MIXED are also given; here, negative changes are good (error reduction).

	Correct		Incorrect			Total
	Cor	Seen	Score	Seen	Sense	
Medical						
Cor	48.3	0.0	**3.1**	0.0	0.0	51.5
Seen-C	**1.6**	2.8	**0.1**	0.0	0.0	4.5
Score	**5.3**	0.0	13.3	0.0	0.0	18.6
Seen-I	**2.3**	0.0	**0.5**	4.0	**2.5**	9.3
Sense	**2.3**	0.0	**2.6**	0.0	11.3	16.2
Total	59.8	2.8	19.6	4.0	13.8	100
Science						
Cor	49.8	0.0	**3.6**	0.0	0.0	53.3
Seen-C	**1.4**	1.4	0.0	0.0	0.0	2.9
Score	**5.8**	0.0	14.2	0.0	0.0	20.0
Seen-I	**1.8**	0.0	**0.3**	5.6	**2.5**	10.2
Sense	**1.4**	0.0	**1.6**	0.0	10.6	13.6
Total	60.1	1.4	19.8	5.6	13.1	100

Table 8.3: Percent of WADE annotation changes moving from OLD (rows) to MIXED (columns) models, for each domain. Non-zero off-diagonals are bolded. Seen-C indicates Freebies, and Seen-I indicates unseen words that were mistranslated.

one estimated over the English side of the Hansard data and the other estimated over the English side of the domain-specific parallel corpus. In later experiments, we also use the English side of the parallel corpora to identify new-domain like comparable corpora (see Section 8.3). We tune the models using batch MIRA (Cherry and Foster, 2012) and development sets in each domain and then use the models to translate each test set.

In our experiments below, we *do not* make use of the French side of the new-domain parallel corpora at all. However, for the purposes of our comparison analyses here, we train a single model on the combination of all old-domain and all new-domain parallel training data. Recall that the key idea behind TETRA is to enhance a baseline translation model, in this case, OLD, an MT system trained on old domain parallel text, to compare the impact of potential sources of improvement. We use parallel new domain data to propose enhancements to the OLD system. This provides a realistic measure of what could be achieved if one had access to parallel data in the new domain. The specific system we build, called MIXED, is a linear interpolation of a translation model trained only on old domain data and a model trained only on new domain data (Foster and Kuhn, 2007). The mixing weights are selected via grid search on a tuning set, selecting for BLEU. We also present WADE analyses on the MIXED system, which, again, provides insight into the potential impact of targeting improvements in each error category.

Table 8.2 shows the results of a WADE analysis on the machine translated test set in each domain. As in Section 3.4, we assume that SEARCH is not a major source of error. In both domains, the WADE analysis shows that just over 50% of the alignment links between test set sentences and their references are translated correctly by the old-domain model, and the percent of total correct alignment links increases by about 5% under MIXED. For both domains, most of the errors analyzed by WADE are SCORE errors, followed by SENSE and then SEEN. The number of SEEN errors goes down dramatically when moving from the OLD model to the MIXED. This is somewhat less true for SENSE errors. At first glance, it appears that the MIXED model is *worse* in terms of SCORE errors according to WADE. However, recall that in our WADE analysis, many errors that used to be SEEN and SENSE errors in the OLD model may become SCORE errors in the new domain.

To see the full picture, we must look at how the different error categories *change* from the OLD system to the

Domain	OLD	OLD +SEEN		OLD +SENSE		OLD +SCORE		MIXED
Medical	28.69	31.02	(+ 8%)	30.59	(+ 7%)	30.21	(+ 5%)	36.60
Science	26.13	27.72	(+ 6%)	27.29	(+ 4%)	28.68	(+10%)	32.23

Table 8.4: TETRA results on science and medical domains using OLD and MIXED models (first and last columns), OLD enhanced with seen translations (second), sense translations (third), and scores (fourth), together with percent improvements in terms of BLEU score. Here, positive improvements are good (higher BLEU scores).

MIXED system in WADE. This is shown in Table 8.3. In this table, the rightmost column contains the total percentage of errors in the OLD systems; the rows labeled *Total* show the total percentage of errors in the MIXED systems; the remaining cells these errors changing from OLD to MIXED. For the medical domain, the OLD system has 18.6% SCORE errors. Of those, 5.3% are fixed in the MIXED system.

For our two domains of interest, addressing SEEN errors can be substantially helpful, in terms of both BLEU score and the fine-grained distinctions considered by WADE. The more interesting conclusion, however, is that simply bringing in new words isn't enough. Table 8.3 shows that in these two domains there are a substantial number of errors that transition from being SEEN-Incorrect to SENSE-Incorrect. This indicates that besides observing a new word, we must also observe it with all of its correct translations.

Likewise, there is a lot to be gained in BLEU by correcting new SENSE translation errors (essentially the same percentage as for SEEN). But this is harder to solve. We can see in Table 8.3 that from the SENSE errors of the OLD system, half become correct but the other half become SCORE errors. So giving appropriate scores to the new senses is a challenge. This makes sense: these new sense are now "competing" with old ones, and getting the interpolation right between old and new domain tables is difficult.

Table 8.4 shows the results of our TETRA analysis, which is largely consistent with what we see in the WADE analysis. That is, for both domains, fixing SEEN errors can be substantially helpful in terms of BLEU score as well as the finer-grained distinctions considered by WADE. The same is true to a lesser degree for SENSE errors. The more interesting conclusion, however, is that simply translating new words is not enough; we must also observe a new word with all of its correct translations and translation scores.

In Irvine *et al.* (2013a), we concluded that that majority of the difference between the performance of the old-domain model and each of the mixed-domain models, which are trained using both old and new domain parallel data, is due to SEEN and SENSE errors. Old-domain models also make many SCORE errors. However, they can be difficult to correct, even given some new domain parallel training data, because manipulating scores can hurt as often as it helps. In our experiments in Section 8.3, rather than manipulate the existing OLD model scores, we *augment* a model with new features estimated using new-domain comparable corpora.

8.3 New-Domain Comparable Corpora

We once again use Wikipedia as a source of comparable corpora. There are over half a million pairs of inter-lingually linked French and English Wikipedia documents.[3] We assume that we have enough monolingual new-domain data in one language to rank Wikipedia pages according to how *new-domain-like* they are. In particular, in our experiments here, we use our new-domain English language modeling data to measure new-domain-likeness. We could have targeted our learning even more by using our new-domain French test sets to select comparable corpora. Doing so may increase the similarity between our test data and comparable corpora. However, to avoid overfitting any particular test set, we use our large English new-domain LM corpus instead.

For each inter-lingually linked pair of French and English Wikipedia documents, we compute the percent of English phrases up to length four that are observed in the English monolingual new-domain corpus and rank document pairs by the geometric mean of the four overlap measures. More sophisticated ways to identify new-domain-like Wikipedia pages (e.g. Moore and Lewis (2010)) may yield additional performance gains, but, qualitatively, the ranked Wikipedia pages seem reasonable for the purposes of generating a large set of top-k new-domain document pairs. The top-10 ranked pages for each domain are listed in Table 8.5. The top ranked science domain pages are primarily related to concepts from the field of physics but also include computer science and chemistry topics. The top ranked medical domain pages are nearly all prescription drugs, which makes sense given the content of the EMEA medical corpus.

[3] We limit our set to those pages that are at least 500 words long, avoiding pages with only small amounts of natural language content.

Science	Medical
Diagnosis (artificial intelligence)	Pregabalin
Absorption spectroscopy	Cetuximab
Spectral line	Fluconazole
Chemical kinetics	Calcitonin
Mahalanobis distance	Pregnancy category
Dynamic light scattering	Trazodone
Amorphous solid	Rivaroxaban
Magnetic hyperthermia	Spironolactone
Photoelasticity	Anakinra
Galaxy rotation curve	Cladribine

Table 8.5: Top 10 Wikipedia articles ranked by their similarity to a new-domain English corpus.

In our experiments in Section 8.4, we explore the effect of varying the number of comparable document pairs used to estimate the new feature scores.

8.4 Using Comparable Corpora to Score Phrase Tables for Domain Adaptation

We begin with a scored phrase table estimated using our old-domain (Hansard) parallel training corpus. Then, we use the methods described in Chapter 5 to supplement the baseline model with additional translation scores estimated over new-domain comparable corpora. We rank English Wikipedia documents according to how new-domain-like they are and use the top-k pages to estimate the following additional phrase table features: (1) phrasal context similarity, (2) lexical context similarity, (3) phrasal topic similarity, (4) lexical topic similarity, and (5) lexical string similarity. The details of each feature function are given in Chapter 5.

These similarity metrics all allow for scores of zero, which can be problematic for our log-linear translation models. We use our second tuning sets[4] to tune a minimum threshold parameter for our new features. We measure performance in terms of BLEU score on the second tuning set as we vary the new feature threshold between $1e-07$ and 0.5 for each domain. A threshold of 0.01, for example, means that we replace all feature with values less than 0.01 with 0.01. For both new-domains, performance drops when we use thresholds lower than 0.01 and higher than 0.25. We use a minimum threshold of 0.1 for all experiments presented below for both domains.

We word align our old-domain training corpus using GIZA++ and use the Moses SMT toolkit (Koehn *et al.*, 2007) to extract a phrase table. Our baseline models use a phrase limit of seven and the standard phrase-based SMT feature set, including forward and backward phrase and lexical translation probabilities. Additionally, we use the standard lexicalized reordering model. We experiment with three 5-gram language models trained using SRILM with Kneser-Ney smoothing on (1) the English side of the Hansard training corpus, (2) the relevant new-domain monolingual English corpus, and (3) the English side of our comparable corpora. We use the top $50,000$ most new-domain-like English Wikipedia documents (about 60 million words for each domain) to train our comparable corpora language models.[5] We experiment with different language model combinations to measure how much performance varies with our access to monolingual new-domain language modeling data.

Our first comparison system augments the standard feature set with the orthographic similarity feature, which is not based on comparable corpora. Our second comparison system uses both the orthographic feature and the contextual and topic similarity features estimated over a *random* set of comparable document pairs. The third system estimates contextual and topic similarity using new-domain-like comparable corpora. We tune our phrase table feature weights for each model separately using batch MIRA (Cherry and Foster, 2012) and new-domain tuning data. Results are averaged over three tuning runs, and we use the implementation of approximate randomization released by Clark *et al.* (2011) to measure whether the output of each feature-augmented model is statistically significantly different from the baseline model that uses the same language model.

[4] *test2* datasets released by Irvine *et al.* (2013a)

[5] We experimented with language models estimated over varying numbers of top-k new-domain-like English documents and did not observe any

Language Model(s)		Medical	Science
Old	Baseline	22.70	21.29
	+ Orthographic Feature	23.09* (+0.4)	21.86* (+0.6)
	+ Orthographic & Random CC Features	23.22* (+0.5)	21.88* (+0.6)
	+ Orthographic & New-domain CC Features	23.98* (+1.3)	22.55* (+1.3)
Old+New	Baseline	28.82	26.18
	+ Orthographic Feature	29.02 (+0.2)	26.40* (+0.2)
	+ Orthographic & Random CC Features	28.86 (+0.0)	26.52* (+0.3)
	+ Orthographic & New-domain CC Features	29.16* (+0.3)	26.50* (+0.3)
Old+CC	Baseline	22.70	21.71
	+ Orthographic Feature	23.14* (+0.4)	22.05* (+0.3)
	+ Orthographic & Random CC Features	23.23* (+0.5)	22.40* (+0.7)
	+ Orthographic & New-domain CC Features	23.62* (+0.9)	22.96* (+1.3)

Table 8.6: Comparison between the performance of baseline old-domain translation models and domain-adapted models in translating science and medical domain text. We experiment with different combinations of three language models: *old*, trained on the English side of our Hansard old-domain training corpus, *new*, trained on the English side of the parallel training data in each new domain, and *cc*, trained on the English side of our comparable corpora. We use comparable corpora of $5,000$ (1) random, and (2) the most new-domain-like document pairs to score phrase tables. All results are averaged over three tuning runs, and we perform statistical significance testing comparing each system augmented with additional features with the baseline system that uses the same LMs. * indicates that the BLEU scores are statistically significantly different to a p-value of less than 0.01 compared with the baselines.

Table 8.6 presents a summary of our results on the test set in each domain.[6] We compare (1) a baseline SMT model, (2) our baseline augmented with features estimated over 5 thousand randomly selected pairs of Wikipedia pages, (3) our baseline augmented with features estimated over 5 thousand pairs of Wikipedia pages selected for their similarity with our new-domain target language corpus.

Using only the old-domain LM, our baselines yield BLEU scores of 22.70 and 21.29 on the medical and science test sets, respectively. When we add the orthographic similarity feature, BLEU scores increase significantly, by about 0.4 on the medical data and 0.6 on science. Adding the contextual and topic features estimated over a random selection of comparable document pairs improves BLEU scores slightly in both domains. Finally, using the most new-domain like document pairs to estimate the contextual and topic features yields a 1.3 BLEU score improvement over the baseline in both domains. For both domains, this result is a statistically significant improvement[7] over each of the first three systems.

In both domains, the new-domain language models contribute substantially to translation quality. Baseline BLEU scores increase by about 6 and 5 BLEU scores points in the medical and science domains, respectively, when we add the new-domain LMs. In the medical domain, neither the orthographic feature nor the orthographic feature in combination with contextual and topic features estimated over random document pairs results in a significant BLEU score improvement. However, using the orthographic feature and the contextual and topic features estimated over new-domain document pairs yields a small but significant improvement of 0.3 BLEU. In the science domain, in contrast, all three augmented models perform statistically significantly better than the baseline. Contextual and topic features yield only a slight improvement above the model that uses only the orthographic feature, but the difference is statistically significant. For the science domain, when we use the new domain language model, there is no difference between estimating the contextual and topic features over random comparable document pairs and those chosen for their similarity with new-domain data.

In the medical domain, the language model trained on $50,000$ English documents in our comparable corpora does not improve performance above the baseline that uses only the old domain LM. In contrast, with the new language model, BLEU scores increase by about 0.4 in the science domain in all experimental conditions. Differences across domains may be due to the fact that the medical domain corpora are much more homogenous, containing the often boilerplate text of prescription drug labels, than the science domain corpora. The science domain corpora, in contrast,

significant difference in performance; none of the LMs improve translation quality.

[6] The slight differences between baseline BLEU scores and those presented in Table 8.4 are due only to variation in tuning. In each table we presented those published in our prior work, Irvine *et al.* (2013a) and Irvine and Callison-Burch (2014b).

[7] p-value < 0.01

	Training Data Word Frequency	Percent of Words Medical	Science
Tokens	0 (OOV)	7.3%	5.8%
Type		23.5%	27.8%
Tokens	$\leqslant 5$	9.6%	7.7%
Type		30.4%	35.3%
Tokens	$\leqslant 10$	10.7%	8.6%
Type		32.9%	38.32%

Table 8.7: Type and token-based OOV rate on the tuning set for each new domain, with respect to the full Hansard corpus.

contain abstracts from several different scientific fields; because that data is more diverse, a randomly chosen mixed-domain set of comparable corpora may still be relevant and useful for adapting a translation model.

Figure 8.1 shows learning curves over a varying number of comparable document pairs for each domain. The simple baseline uses only the standard bilingual phrase table features estimated over the old-domain parallel training corpus and both the old and new LMs. Our proposed approach orders comparable document pairs by how new-domain-like they are and augments models with the orthographic feature as well as the contextual and topic features estimated over the top-k document pairs. As a result, using more comparable document pairs means that there is more data from which to estimate signals, but it also means that the data is less new-domain like overall.

In the medical domain, we do not observe additional performance gains by using more than just a few thousand comparable document pairs to estimate the new features. In fact, performance drops when $10,000$ or more document pairs are used. Again, we attribute this to the homogeneity of the medical data; only a small set of documents in our comparable corpora are in the same domain of text. In contrast, BLEU scores improve very slightly as we add more comparable document pairs in the science domain. These gains are not statistically significant, however.

The two major findings in the experiments presented here are as follows: (1) Doing domain adaptation by using new-domain comparable corpora to score a phrase table estimated over old-domain data can significantly improve translation quality, and (2) Results are highly dependent on the type and size of the initial mixed-domain comparable corpus as well as the homogeneity of the text domain of interest.

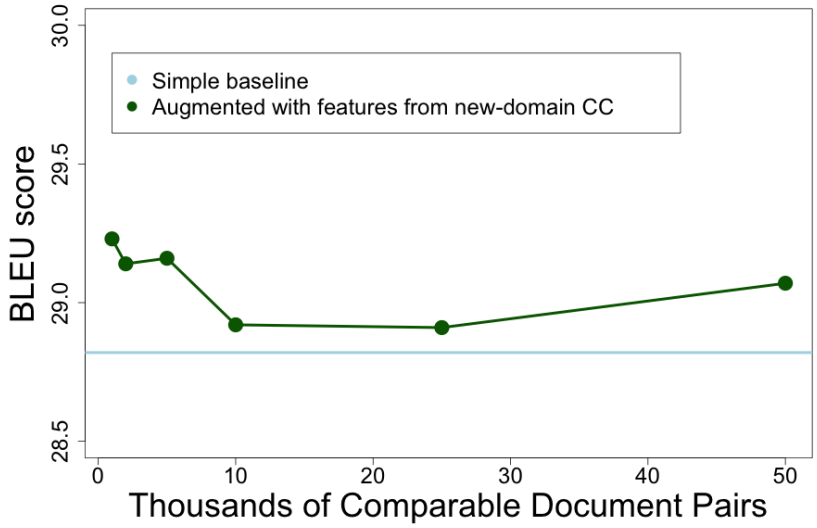
8.5 Using Comparable Corpora to Translate Unseen Words for Domain Adaptation

In Section 8.4, we used new-domain comparable corpora to reduce the number of SCORE errors in a domain adaptation SMT setting. Here, we use comparable corpora to reduce the number of SEEN errors. As was the case for low resource SMT, there is a lot to be gained in terms of translation quality from identifying translations for previously unseen, or OOV, words. As in Chapter 6, we use the bilingual lexicon induction technique introduced in Chapter 4 to identify new unigram translations, in this case for unknown and low frequency words in our *new-domain* tuning and test sets. Table 8.7 gives the rate of OOV and low frequency words for each domain's tuning set with respect to the full Hansard corpus.

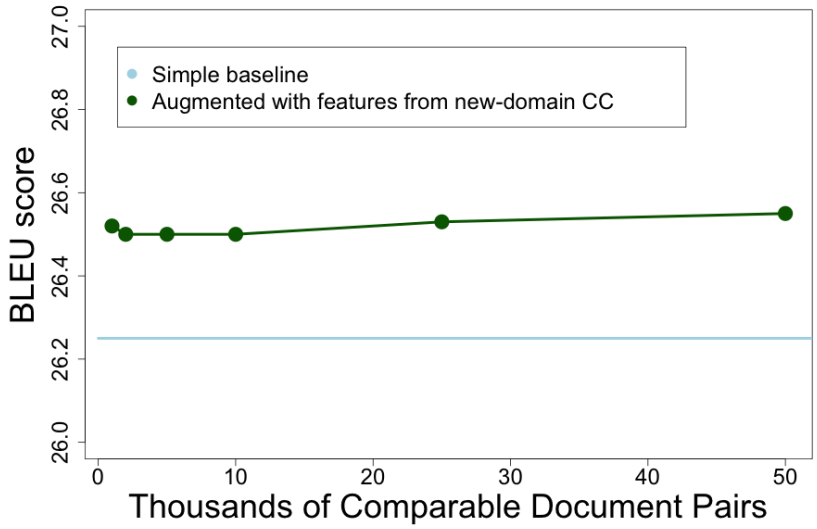
8.5.1 Bilingual Lexicon Induction Model

We train a bilingual lexicon induction model on $3,000$ unigram translations taken from the word-aligned Hansard old-domain corpus. As in Chapter 4, we use three times as many randomly chosen word pairs as examples of non-translations. Here, we only use features extracted from Wikipedia data. Our feature set includes the following, a subset of those used in Chapter 4:

1. Orthographic Similarity
2. Wikipedia Contextual Similarity
3. Wikipedia Topic Similarity
4. Wikipedia Frequency Similarity
5. Wikipedia IDF Similarity
6. Wikipedia Burstiness Similarity



(a) Medical



(b) Science

Figure 8.1: Translation performance in terms of BLEU score as we vary the number of comparable document pairs used for scoring. Our approach ranks comparable document pairs by how new-domain-like they are, so, as the number increases, we estimate features over a larger amount of data, but the data is less homogeneous.

7. String Identity
8. Inverse Log of Target Wikipedia Frequency

We test our model on a held-out set of $2,000$ translations from our old domain data. Our French BLI models achieve top-1 and top-10 accuracies of 44% and 63%, respectively.

8.5.2 Evaluation of Induced Translations

We identify all new-domain words that appear ten or fewer times in our old-domain training data. Using the model learned over old-domain word pairs, we induce a top-k list of translations for each word in our new-domain list. We allow all English words that appear three or more times in our comparable corpora as candidate translations. As usual,

Domain	Comparable Corpus	Top-1 Accuracy	Top-10 Accuracy
Medical	Random 5k	4.1	6.0
	Top 5k Medical	29.7	42.5
	Top 50k Medical	27.5	42.5
Science	Random 5k	6.0	9.0
	Top 5k Science	21.0	31.6
	Top 50k Science	24.9	38.9

Table 8.8: Performance of bilingual lexicon induction on tuning set words in each domain that appear 10 or fewer times in the old-domain training data. Accuracies give the percent of source words for which a correct target language translation appears in its top-k ranked list of induced translations.

before moving to end-to-end SMT experiments, we measure the quality of our induced translations intrinsically. Table 8.8 shows the top-1 and top-10 accuracies of induced translations for the OOV and low frequency words in each domain's tuning set using several different comparable corpora. Overall, performance is substantially lower than what we observed over our held out old-domain test set. However, here, the unigrams for which we would like to induce translations are low frequency. As we showed in Chapter 4, the performance of our bilingual lexicon induction models tends to be lower on lower frequency words.

In Section 8.4, we compared the impact of augmenting a baseline SMT model with new features estimated over a random set of comparable document pairs and those estimated over new-domain comparable document pairs. We perform similar experiments here and compare the accuracy of induced unigram translations when we estimate features over (1) a general corpus of comparable document pairs, and (2) new-domain comparable document pairs. In all experiments, we use the same trained model for ranking candidate translations and only vary how we estimate features over our *test* sets. Because we have assumed that we do not have access to a new-domain parallel training corpus or dictionary, we do not train a domain-specific model of bilingual lexicon induction. For both the medical and the science domain, estimating features over the $5,000$ most new-domain document pairs dramatically outperforms doing so over a random set of $5,000$ document pairs. In the medical domain, top-1 accuracy increases from 4% to 30% when we move from random documents to new-domain document pairs, and in the science domain it increases from 6% to 21%. When we use the $50,000$ instead of the $5,000$ most new-domain-like document pairs to estimate features, we observe only slight performance increases for science and no additional gains on the medical data. This is consistent with our experiments estimating new phrase table features; using more data is slightly helpful in the science domain but not in the medical domain. Again, we attribute this to the relative heterogeneity of the science domain corpora in comparison with the medical domain corpora.

8.5.3 Integrating Translations into End-to-End SMT

Table 8.9 shows the results of our end-to-end SMT experiments where we do domain adaptation by (1) adding new features estimated over new-domain comparable corpora (Section 8.4), improving SCORE errors, (2) adding new translations induced from new-domain comparable corpora, improving SEEN errors, and (3) adding both new scores and new translations. As before, we compare the impact of estimating features and identifying new translations using a random set of comparable document pairs with using new-domain comparable document pairs. We present results using the old-domain language model only as well as both the old and new domain language models. In all experiments, we add the top-1 ranked translation for each OOV and low frequency word as well as accent-stripped identity translations, which we showed to be useful in our prior work (Irvine *et al.*, 2013b).

In general, the results show that augmenting models with translations of OOV and low frequency words does not improve performance in most cases. The one exception is in the medical domain using both the old and new domain LMs; BLEU scores increase by about 0.2 in comparison with the model augmented with scores alone. We attribute this negative result to two things. First, because our old-domain Hansard training corpus is so large, OOV rates are fairly low even in our domain adaptation setting. As Table 8.7 shows, only about 7% and 6% of word tokens in the medical and science tuning sets, respectively, are OOV, which corresponds to about two words per sentence in the science data and about one word per sentence in the medical data.[8] Secondly, our top-1 accuracies are only about 30% and 20% on

[8]Sentences in the science data are much longer, about 32 words on average, than those in the medical data, which are about 16 words long on

Domain	Comparable Corpus	SMT Model	Language Model: Old	Language Model: Old + New
Medical	Baseline		22.70	28.82
	Random 5k	+ Scores	23.22* (+0.5)	28.86 (+0.0)
		+ Induced Translations	22.12* (−0.6)	28.27* (−0.6)
		+ Scores & Induced Translations	22.46 (−0.2)	28.26* (−0.6)
	Top 5k Medical	+ Scores	**23.98*** (+1.3)	29.16* (+0.3)
		+ Induced Translations	22.32 (−0.4)	28.88 (+0.1)
		+ Scores & Induced Translations	23.62* (+0.9)	**29.35*** (+0.5)
Science	Baseline		21.29	26.18
	Random 5k	+ Scores	21.88* (+0.6)	**26.52** (+0.3)
		+ Induced Translations	20.20* (−1.1)	25.39* (−1.1)
		+ Scores & Induced Translations	21.02* (−0.3)	25.69* (−0.8)
	Top 5k Science	+ Scores	**22.55*** (+1.3)	26.50 (+0.3)
		+ Induced Translations	20.39* (−0.9)	25.90* (−0.3)
		+ Scores & Induced Translations	21.45 (+0.2)	26.29 (+0.1)

Table 8.9: BLEU score results applying our methods for scoring a phrase table (+ Scores), inducing translations for OOV and low frequency words (+ Induced Translations), and both (+ Scores & Induced Translations) in a domain adaptation SMT setting. The highest BLEU scores for each domain and LM condition are highlighted.

the medical and science data, respectively. Because OOV rates are low and our lexicon induction accuracy is modest, our efforts to improve SEEN errors yield small or no gains. Although there are more low frequency words than OOV words, unlike the low resource language pair setting that we explored in Chapter 6, because our old-domain baselines are trained on such a large corpus, there are likely to be many fewer alignment errors, and, hence, our models are more likely to contain accurate translations for low frequency items.

In our prior work in Irvine *et al.* (2013b), we used an alternate approach to inducing translations for OOV and low frequency words in a domain adaptation setting and saw substantial gains in BLEU scores. However, in that work, we limited our old-domain training set to every 32nd line in the original set, making the baseline model much weaker and easier to improve. We would likely see more gains from augmenting models with induced translations using the methods presented here if our initial baselines were also weaker.

8.5.4 Conclusion

In this chapter, we applied our methods for using comparable corpora to score an existing phrase table and induce translations to the domain adaptation setting. We showed that using comparable corpora selected for their new-domain-likeness to score a phrase table resulted in statistically significant improvements in French-English translation quality above very strong baseline systems for both the science and medical domains. We also presented one negative result; using our methods for inducing translations for OOV words and augmenting models in the same domain adaptation setting did not result in improvements in translation quality. We attribute this primarily to the strength of our baseline models, which have low OOV rates.

average.

Chapter 9

Conclusion

The performance of statistical machine translation models is heavily dependent upon the amount and type of parallel data used for training. For some language pairs and domains of text, for example French-English newswire or parliamentary proceedings, very large amounts of parallel data are available and the quality of machine translation has improved dramatically in the past two decades as methods for learning models from such data have matured. However, for many language pairs and domains of text, very little or no parallel data is available. Without parallel data, we are unable to use standard methods to estimate high quality statistical translation model parameters and, thus, are unable to generate high quality machine translations. In this thesis, we used *comparable corpora* to improve the performance of SMT models trained using little or no parallel data. In particular, in this thesis we used comparable corpora to identify new word and phrase translations and to estimate translation probabilities. Many of our techniques were taken from prior work in bilingual lexicon induction; one of the major contributions of this thesis is the application of such techniques to end-to-end SMT.

In Chapter 3 we presented dictionaries and comparable corpora that we have collected for over 150 human languages. Our dictionaries come from a variety of sources, including scanned paper dictionaries and through crowdsourcing efforts. Although some dictionaries were publicly available already, others were not, and the compilation of new and existing dictionaries is a valuable resource for ongoing research in multilingual natural language processing. Our comparable corpora come from Wikipedia and crawls of online news websites. Wikipedia is freely available for download and has been used extensively in prior NLP research. One small contribution of this thesis is the release of Wikipedia data for 142 foreign languages which has already been extracted from the original source HTML, preprocessed, and paired with inter-lingually linked English pages. The comparable news data that we release is a result of a multi-year web crawl effort. Unlike the Wikipedia data, our news crawl data is a novel resource. Like the Wikipedia data, we release preprocessed datasets for over 100 languages.

This thesis presents two novel error analysis techniques and makes use of both throughout. In Chapter 3, we presented Table Enhancement for Translation Analysis (TETRA) and Word Alignment Driven Evaluation (WADE), which are two methods for measuring different types of lexical choice errors. In Irvine *et al.* (2013a), we released code for running both types of analysis. In Section 6.4.4.2, we expanded our original WADE definition to take advantage of multiple reference translations. Sections 3.4.2 and 8.2 gave the results of our analyses over outputs produced by low resource models trained on small amounts of parallel data and those produced by models trained on old-domain data, respectively. In both settings, the results of our analyses showed the major sources of error result from unseen source language words and unseen translations. In Sections 6.4.4.2 and 8.2, we showed that augmenting models with new translations is not enough; we must also score them accurately.

In Chapter 4 we presented a new supervised discriminative approach to inducing word translations. We defined and used a variety of features estimated over comparable corpora and a small number of example translations to learn a model for classifying word pairs as translations or non-translations. We provided an extensive analysis of the effectiveness of the following signals, most of which we estimated using comparable corpora: contextual, temporal, orthographic, topic, frequency, and burstiness similarity. Most features had been used in prior work on bilingual lexicon induction. The one exception is our topic feature, which we defined and which proved to be one of the most valuable signals for predicting translations. Using a diverse feature set, we observed gains on the bilingual lexicon induction task of greater than 100% over an unsupervised rank-combination baseline using the same feature set. We also observed gains of greater than 60% over a previously state-of-the-art model (Haghighi *et al.*, 2008) using the same feature subset

as the prior work. When we used our full feature set, we saw improvements of over 180% above the Haghighi *et al.* (2008) baseline. We showed that our proposed model is robust to the amount and type of training data; high quality models may even be learned using supervision from a different language pair. Additional experiments showed that bursty, frequent words are easier to translate than less bursty, infrequent words. Throughout our bilingual lexicon induction experiments, we presented results inducing English translations for source words in 24 foreign languages, emphasizing the general applicability of our approach.

In Chapter 5, we adapted our techniques for estimating signals of translation equivalence for *pairs of words* in order to score *pairs of phrases*. In our experiments, we showed that, given a high-quality but noisy phrase table, our new comparable corpora-based feature functions did a good job of distinguishing high quality from low quality phrase pairs. When we dropped all of the bilingually estimated features from a phrase table and replaced them with our monolingual equivalents, we were able to regain 56% of the loss in our Spanish-English setting and 36% of the loss in our Urdu-English experiments. These experiments demonstrated the strength of the features estimated over comparable corpora.

In Chapter 6, we dropped the Chapter 5 assumption that we were given a set of high quality phrase pairs. Instead, we assumed a more realistic setting in which we had access to a small amount of parallel text from which to estimate a baseline SMT model. We augmented the low resource SMT models with both new translations identified using our discriminative bilingual lexicon induction model (Chapter 4) and new phrase pair features estimated over comparable corpora (Chapter 5). We found that improvements resulting from new translations and new features are nearly additive, and we observed total BLEU score gains of up to 1.5 points for the following truly low resource languages: Bengali, Hindi, Malayalam, Tamil, Telugu, and Urdu. Although the gains in BLEU score were substantial, the gains in terms of the readability of the output machine translations were dramatic. This was largely due to our models translating previously unknown words. Additional experiments showed that each of the following can also improve performance: augmenting models with more than a single translation for source words, inducing translations for low frequency source words as well as OOV words, and using as much comparable corpora as possible to estimate new feature functions.

Chapter 6 mainly presents experimental results building upon the novel ideas developed in Chapters 4 and 5. However, the Chapter 6 experiments are a valuable result in that they bridge prior work on bilingual lexicon induction with end-to-end machine translation. Although prior work on bilingual lexicon induction has typically pointed to machine translation as an application, very little prior work has actually incorporated induced bilingual lexicons into SMT models. In fact, a lot of prior work only induced translations for frequent nouns (e.g. Haghighi *et al.* (2008)), and it remained to be seen if inducing translations for, for example, OOV words and augmenting SMT models with the new phrase pairs would improve translation quality. We showed that induced translations integrated into end-to-end SMT can improve machine translations and measured their effect in a variety of experimental conditions. Also unlike prior work, we found it important and valuable to experiment with truly low resource languages.

In Chapter 7, we proposed a method for *composing* phrase translations from multiple unigram translations. The composition algorithm that we defined and used is more flexible than prior definitions of compositionality. That is, it generates more translation candidates for source phrases than more strict definitions. We provided a rigorous analysis of the tradeoffs between generating high-precision, lower recall sets of composed phrase translations and generating low-precision, higher recall sets. We deliberately chose a fairly high-recall algorithm for composing translations and then used features estimated over comparable corpora to prune and rank large sets of hypothesis composed phrase translations. For both Spanish-English and Hindi-English translation, we observed improvements in translation quality of over half a BLEU point above baselines augmented with induced unigram translations and monolingual features.

In Chapter 8 we used comparable corpora to tackle the domain shift problem in machine translation, which is closely related to that of working with low resource language pairs. In this setting, we have large amounts of parallel training data in some old domain of text but not enough training data in the new domain to train a high quality SMT model. We experimented with French to English translation of medical and scientific text given large amounts of parliamentary proceedings old-domain parallel text because large, public datasets exist for that language pair and those domains. We used our methods for inducing new translations and scoring existing phrase pairs to augment baseline models and observed gains in translation quality of up to 1.3 BLEU points over very strong French-English baselines. As was the case for low resource language pairs, our approaches improve translation quality more than BLEU scores may suggest because we target and translate previously unknown words, which contribute substantially to readability. Like the results in Chapter 6, Chapter 8 largely presents experimental results employing the techniques developed in previous chapters. However, these experimental results provide valuable insight into the potential for using comparable corpora to augment SMT models in a variety of settings. In particular, like Chapter 6, Chapter 8 presents novel results applying bilingual lexicon induction to end-to-end SMT.

We believe that making use of alternative data resources is critical for expanding the applicability of statistical machine translation beyond the few language pairs and domains for which large parallel datasets exist. In this thesis, we have taken advantage of one such data resource, *comparable corpora*. In particular, we have improved and applied previous approaches to bilingual lexicon induction to end-to-end SMT. Our experiments show that augmenting baseline SMT systems with new translations and features estimated over comparable corpora can improve translation performance significantly for truly low resource language pairs.

Chapter 10

Language Set

Language	Family	Location	Word Order	Affixing
Afrikaans	Indo-European	South Africa	-	-
Albanian	Indo-European	Albania, Serbia, Montenegro	SVO	Strongly suffixing
Alemannic German	Indo-European	Switzerland, Germany, Austria, etc.	None	Strongly suffixing
Amharic	Afro-Asiatic	Ethiopia	SOV	Weakly suffixing
Ao	Sino-Tibetan	India	SOV	Weakly suffixing
Arabic	Afro-Asiatic	Africa, Middle East, etc.	VSO	Strongly suffixing
Aragonese	Indo-European	Spain	-	-
Armenian	Indo-European	Armenia	None	Strongly suffixing
Assamese	Indo-European	India	SOV	Strongly suffixing
Asturian	Indo-European	Spain	-	-
Azeri	Altaic	Azerbaijani	SOV	Strongly suffixing
Bashkir	Altaic	Russia	SOV	Strongly suffixing
Basque	Basque	France, Spain	SOV	Equal
Belarusian	Indo-European	Belarus	None	Strongly suffixing
Bengali	Indo-European	Bangladesh, India	SOV	-
Bihari	Indo-European	India	-	-
Bishnupriya Manipuri	Indo-European	India, Bangladesh, Myanmar	-	-
Bosnian	Indo-European	Bosnia-Herzegovina	-	-
Breton	Indo-European	France	SVO	Weakly suffixing
Buginese	Austronesian	Indonesia	SVO	-
Bulgarian	Indo-European	Bulgaria	SVO	Strongly suffixing
Burmese	Sino-Tibetan	Myanmar	SOV	Strongly suffixing
Catalan	Indo-European	Spain	SVO	-
Cebuano	Austronesian	Philippines	VSO	-
Central Bicolano	Austronesian	Philippines	-	Little affixation
Chinese (Mandarin)	China	Sino-Tibetan	SVO	Strongly suffixing
Chuvash	Altaic	Russia	SOV	Strongly suffixing
Croatian	Indo-European	Croatia	SVO	Strongly suffixing
Czech	Indo-European	Czech Republic	SVO	Weakly suffixing
Danish	Indo-European	Denmark	SVO	Strongly suffixing
Dari	Indo-European	Afghanistan	-	-
Dhivehi	Indo-European	Maldives	-	-
Dutch	Indo-European	Netherlands	None	Strongly suffixing
Esperanto	Constructed	-	-	-
Estonian	Uralic	Estonia	SVO	Strongly suffixing
Faroese	Indo-European	Denmark	-	-
Farsi	Indo-European	Iran	SOV	Weakly suffixing
Finnish	Uralic	Finland	SVO	Strongly suffixing
French	Indo-European	France, Switzerland	SVO	Strongly suffixing
Galician	Indo-European	Spain	-	-
Georgian	Kartvelian	Georgia	SOV	Weakly suffixing
German	Indo-European	Austria, Germany, Switzerland	None	Strongly suffixing
Gilbertese (Kiribati)	Austronesian	Kiribati	VOS	Equal
Goan Konkani	Indo-European	India	-	-
Greek	Indo-European	Greece	None	Strongly suffixing
Gujarati	Indo-European	India	SOV	-

Table 10.1 – continued from previous page

Language	Family	Location	Word Order	Affixing
Haitian	Other (Creole)	Haiti	-	-
Hebrew	Afro-Asiatic	Israel	SVO	Weakly suffixing
Hindi	Indo-European	India	SOV	Strongly suffixing
Hungarian	Uralic	Hungary	None	Strongly suffixing
Icelandic	Indo-European	Iceland	SVO	Strongly suffixing
Ido	Constructed	-	-	-
Ilokano	Austronesian	Philippines	VSO	-
Indonesian	Austronesian	Indonesia	SVO	Strongly suffixing
Irish	Indo-European	Ireland	VSO	Equal
Italian	Indo-European	Italy, Switzerland	SVO	Strongly suffixing
Japanese	Japanese	Japan	SOV	Strongly suffixing
Javanese	Austronesian	Indonesia	-	-
Kalaallisut (West Greenlandic)	Eskimo-Alteut	Greenland	SOV	Strongly suffixing
Kannada	Dravidian	India	SOV	Strongly suffixing
Kapampangan	Austronesian	Philippines	-	-
Kashmiri	Indo-European	India, Pakistan	SVO	-
Kazakh	Altaic	Kazakhstan	-	-
Khasi	Austro-Asiatic	India	SVO	Little affixation
Khmer	Austro-Asiatic	Cambodia	SVO	Little affixation
Kinyarwanda	Niger-Congo	Rwanda	SVO	-
Korean	Korean	North Korea, South Korea	SOV	Strongly suffixing
Kurdish	Indo-European	Iran, Iraq	SOV	Weakly suffixing
Kyrgyz	Altaic	Kyrgyzstan	-	Strongly suffixing
Lao	Tai-Kadai	Laos, Thailand	SVO	Little affixation
Latin	Indo-European	-	-	-
Latvian	Indo-European	Latvia	SVO	Weakly suffixing
Lithuanian	Indo-European	Lithuania	SVO	Strongly suffixing
Lombard	Indo-European	Italy, Switzerland	-	-
Low Saxon	Indo-European	Germany	-	-
Luganda	Niger-Congo	Uganda	-	Strongly prefixing
Luxembourgish	Indo-European	Luxembourg	-	-
Macedonian	Indo-European	Macedonia	SVO	-
Malagasy	Austronesian	Madagascar	VOS	Little affixation
Malayalam	Dravidian	India	SOV	-
Malaysian	Austronesian	Malaysia	-	-
Marathi	Indo-European	India	SOV	Strongly suffixing
Maltese	Afro-Asiatic	Malta	-	-
Maori	Austronesian	New Zealand	VSO	Little affixation
Mizo	Sino-Tibetan	Bangladesh, India	SOV	Little affixation
Mongolian (Khalkha)	Altaic	Mongolia	SOV	Strongly suffixing
Montenegrin	Indo-European	Montenegro	SVO	Strongly suffixing
Neapolitan	Indo-European	Italy	-	-
Nepali	Indo-European	Nepal	SOV	Strongly suffixing
Newar / Napal Bhasa	Sino-Tibetan	Nepal	SOV	Strongly suffixing
Norwegian	Indo-European	Norway	SVO	Strongly suffixing
Norwegian (nynorsk)	Indo-European	Norway	SVO	Strongly suffixing
Occitan	Indo-European	France	-	-
Oriya	Indo-European	India	SOV	Strongly suffixing
Pangasinan	Austronesian	Philippines	VSO	-
Panjabi	Indo-European	India, Pakistan	SOV	Strongly suffixing
Papiamento	Other (Creole)	Netherlands, Antilles	-	-
Pashto	Indo-European	Afghanistan, Pakistan	SOV	Strongly suffixing
Piedmontese	Indo-European	Italy	-	-
Polish	Indo-European	Poland	SVO	Strongly suffixing
Portuguese	Indo-European	Portugal	SVO	Strongly suffixing
Quechua	Quechuan	Peru, Bolivia, Equador	SOV	Strongly suffixing
Romani	Indo-European	United Kingdom	None	Strongly suffixing
Romanian	Indo-European	Romania	SVO	Strongly suffixing
Romansh	Indo-European	Switzerland	-	-
Russian	Indo-European	Russia	SVO	Strongly suffixing
Sami	Uralic	Finland, Norway, Sweden	SVO	Strongly suffixing
Sanskrit	Indo-European	India	-	-
Serbian	Indo-European	Serbia	SVO	Strongly suffixing
Serbo-Croatian	Indo-European	Bosnia-Herzegovina, Croatia, Serbia, and Montenegro	SVO	Strongly suffixing

Table 10.1 – continued from previous page

Language	Family	Location	Word Order	Affixing
Shona	Niger-Congo	Zimbabwe	SVO	Strongly prefixing
Sicilian	Indo-European	Italy	-	-
Sindhi	Indo-European	India, Pakistan	-	-
Sinhalese	Indo-European	Sri Lanka	SOV	-
Slovak	Indo-European	Slovakia	-	-
Slovenian	Indo-European	Slovenia	SVO	Strongly suffixing
Somali	Afro-Asiatic	Somalia	SOV	Strongly suffixing
Spanish	Indo-European	Spain	SVO	Strongly suffixing
Sundanese	Austronesian	Indonesia	SVO	Little affixation
Swahili	Niger-Congo	Tanzania	SVO	Weakly prefixing
Swedish	Indo-European	Finland, Sweden	SVO	Strongly suffixing
Tagalog	Austronesian	Philippines	VSO	Little affixation
Tajik	Indo-European	Tajikistan	SOV	Weakly suffixing
Tamil	Dravidian	India, Sri Lanka	SOV	Strongly suffixing
Tatar	Altaic	Russia	SOV	Strongly suffixing
Telugu	Dravidian	India	SOV	Strongly suffixing
Tetum	Austronesian	East Timor	SVO	Little affixation
Thai	Tai-Kadai	Thailand	SVO	Little affixation
Tibetan	Sino-Tibetan	China	SOV	-
Tigre	Afro-Asiatic	Eritrea	SOV	Weakly suffixing
Tigrinya	Afro-Asiatic	Eritrea, Ethiopia	SOV	-
Tongan	Austronesian	Tonga	VSO/VOS	Little affixation
Tok Pisin	Other (Pidgin)	Papua New Guinea	-	-
Turkish	Altaic	Turkey	SOV	Strongly suffixing
Turkmen	Altaic	Turkmenistan	SOV	-
Uighur	Altaic	China	SOV	Strongly suffixing
Ukrainian	Indo-European	Ukraine	SVO	Strongly suffixing
Urdu	Indo-European	Pakistan	SOV	Strongly suffixing
Uzbek	Altaic	Afghanistan, Uzbekistan	SOV	Strongly suffixing
Valencian	Indo-European	Spain	-	-
Vietnamese	Austro-Asiatic	Vietnam	SVO	Little affixation
Volapk	Constructed	-	-	-
Walloon	Indo-European	Belgium, France	-	-
Waray-Waray	Austronesian	Philippines	-	-
Welsh	Indo-European	United Kingdom	VSO	Strongly suffixing
Western Panjabi	Indo-European	India, Pakistan	-	-
West Frisian	Indo-European	Netherlands	-	-
Wolof	Niger-Congo	Gambia, Senegal	SVO	Little affixation
Yoruba	Niger-Congo	Benin, Nigeria	SVO	Little affixation
Zazaki	Indo-European	Turkey	SOV	-
Zulu	Niger-Congo	South Africa	SVO	Strongly prefixing

Table 10.1: List of 151 languages for which we release language packs, with linguistic annotations taken from WALS. Blanks indicate missing WALS feature values. Locations are places where a given language is spoken canonically. In some cases, a given language is spoken much more widely than indicated, for example Spanish is spoken not only in Spain but throughout Latin America. If there is no dominant word order, ”None” is listed.

Chapter 11

Data Resources

Language	Thousands of Native Speakers	Thousands of Parallel Tokens
Afrikaans	4949	700
Albanian	7436	21240
Amharic	21811	232
Arabic	223010	452840
Armenian	5924	9
Assamese	12828	21
Asturian	110	100
Basque	657	1502
Belarusian	7818	313
Bengali	193263	1579
Berber	10000	200
Bosnian	2216	14416
Breton	1200	300
Bulgarian	6795	143589
Catalan	7220	2391
Chinese	1197392	1156726
Crimean Tatar	475	10
Croatian	5533	110104
Czech	9469	229625
Danish	5592	206182
Dutch	22984	320001
Estonian	1078	72515
Farsi	56645	8907
Finnish	4994	161035
French	68458	1725721
Gaelic	63	2
Galician	3185	1119
Georgian	4237	88
German	83812	353575
Greek	13068	247447
Gujarati	46633	400
Hausa	24988	3
Hebrew	5302	88383
Hindi	260302	1816
Hungarian	12319	159166
Icelandic	243	7456
Ido	0	10
Indonesian	23200	8281
Irish	106	3401
Italian	61068	318102

Table 11.1 – continued from previous page

Language	Thousands of Native Speakers	Thousands of Parallel Tokens
Japanese	122072	15312
Kalmyk	360	1
Kannada	37739	371
Kashubian	100	100
Kazakh	8077	308
Khmer	14224	800
Kinyarwanda	7189	600
Korean	66418	7224
Kurdish	29960	83
Latvian	1472	37030
Lithuanian	3130	38621
Luxembourgish	320	30
Macedonian	1710	11151
Maithili	32800	100
Malay	59418	17500
Malayalam	33534	774
Maltese	429	19701
Marathi	71780	62
Mongolian	5753	1
Nepali	14410	421
Northern Sami	20	200
Norwegian	4741	35918
Occitan	2048	103
Oriya	50137	29
Panjabi	93171	500
Pashto	26940	208
Pedi	4101	1
Polish	39042	105115
Portuguese	202468	297147
Pushto	26940	8
Quechua	9062	1
Romanian	23623	218135
Russian	161727	207119
Serbian	9262	92433
Sinhala	15577	896
Slovak	5007	56337
Slovene	1906	104119
Somali	16559	318
Spanish	405638	779585
Swedish	8381	144733
Tagalog	24216	23
Tajik	4479	400
Tamil	68763	2167
Tatar	5407	1
Telugu	74049	1353
Thai	20421	6600
Turkish	50733	128300
Uighur	8791	17
Ukrainian	36028	2930
Upper Sorbian	13	88
Urdu	63431	3846
Uzbek	21930	100
Vietnamese	67762	5815
Walloon	600	200
Welsh	536	357
Western Frisian	467	200

Table 11.1 – continued from previous page

Language	Thousands of Native Speakers	Thousands of Parallel Tokens
Xhosa	7817	300

Table 11.1: Thousands of native speakers and thousands of parallel text (paired with English) for each of 96 languages.

Language	Scr	Scanned Paper		Electronic		Crowdsourced		**All Dicts**	
		Source Words	Total Trans.	Source Words	Total Trans.	Source Words	Total Trans.	Source Words	Total Trans.
Afrikaans	r	-	-	-	-	8,223	8,247	8,223	8,247
Albanian	r	-	-	44,245	104,608	9,109	9,127	51,351	112,241
Amharic	o					8,051	8,093	8,051	8,093
Arabic	a	-	-	75,911	149,054	10,103	10,144	83,138	157,886
Aragonese	r	-	-	-	-	2,502	2,518	2,502	2,518
Armenian	o	-	-	-	-	1,806	1,876	1,806	1,876
Asturian	r	-	-	-	-	6,202	6,214	6,202	6,214
Azeri	r	2,047	2,263	225,756	225,808	9,356	9,367	236,442	237,051
Basque	r	-	-	828	880	9,017	9,031	9,440	9,726
Belarusian	c	-	-	-	-	9,884	9,894	9,884	9,894
Bengali	r	-	-	1,073	1,161	-	-	1,073	1,161
	o	-	-	408	444	10,127	10,155	10,364	10,494
Bosnian	r	9,643	15,093	-	-	9,241	9,305	17,428	23,388
Breton	r	-	-	-	-	2,086	2,089	2,086	2,089
Bulgarian	r	-	-	5,057	7,514	-	-	5,057	7,514
	c	-	-	67,663	152,016	10,241	10,287	77,904	162,303
Catalan	r	-	-	-	-	8,827	8,852	8,827	8,852
Cebuano	r	-	-	-	-	6,338	6,637	6,338	6,637
Chinese	o	-	-	54,164	82,080	3,512	3,963	57,294	85,833
Czech	r	-	-	69,659	180,496	9315	9329	75,788	187,474
Danish	r	-	-	-	-	8,092	8,101	8,092	8,101
Dutch	r	-	-	54,519	122,323	7,797	8,120	58,141	127,513
Farsi	r	5,412	6,663	-	-	-	-	5,412	6,663
	a	-	-	96,208	172,722	2,907	2,926	98,448	175,160
Finnish	r	-	-	-	-	8,535	8,550	8,535	8,550
French	r	-	-	45,041	103,267	7,792	8,299	49,030	108,625
Galician	r	-	-	-	-	8,726	8,738	8,726	8,738
Georgian	o	-	-	-	-	5,837	5,897	5,837	5,897
German	r	-	-	54,039	140,818	8,020	8,093	58,015	146,135
Greek	o	-	-	42,921	82,069	10,173	10,185	53,094	92,254
Gujarati	o	-	-	-	-	9,977	19,795	9,977	19,795
Haitian	r	-	-	-	-	5,620	5,663	5,620	5,663
Hebrew	o	-	-	-	-	8,460	8,491	8,460	8,491
Hindi	r	-	-	25,305	58,179	-	-	25,305	58,179
	o	-	-	-	-	9,681	9,824	9,681	9,824
Hungarian	r	-	-	138,585	258,304	8,679	8,771	143,407	264,519
Icelandic	r	-	-	-	-	7,172	7,177	7,172	7,177
Ilokano	r	-	-	-	-	4,318	4,411	4,318	4,411
Indonesian	r	31,799	64,900	1,388	2,143	7,383	7,440	35,442	71,685
Irish	r	-	-	831	887	8,369	8,373	8,769	9,012
Italian	r	-	-	36,663	84,951	7,964	8,056	40,569	89,616
Japanese	o	-	-	-	-	8,366	8,516	8,366	8,516
Javanese	r	-	-	-	-	7,138	7,358	7,138	7,358
Kannada	o	-	-	-	-	10,017	10,192	10,017	10,192
Kapampangan	r	-	-	911	1,000	2,036	2,092	2,750	2,942
Kazakh	r	1,979	2,261	137,403	138,300	30	30	139,303	140,526
Korean	o	-	-	66,384	88,717	7,486	7,506	72,279	95,676
Kurdish	r	-	-	4,159	6,632	33	35	4,186	6,665
Kyrgyz	r	1,897	2,164	67,727	67,773	-	-	69,547	69,890
Latin	r	-	-	6,853	18,884	-	-	6,853	18,884
Latvian	r	-	-	33,486	74,936	9,722	9,746	40,647	83,024
Lithuanian	r	-	-	-	-	9,692	9,706	9,692	9,706
Low Saxon	r	-	-	-	-	5,600	5,651	5,600	5,651
Luxembourgish	r	-	-	-	-	4,610	4,617	4,610	4,617
Macedonian	c	-	-	-	-	9,961	9,968	9,961	9,968
Malagasy	r	-	-	-	-	159	159	159	159
Malayalam	o	-	-	-	-	10,127	10,199	10,127	10,199
Malaysian	r	-	-	5,986	9,438	7,661	7,704	11,292	15,851
Marathi	o	-	-	-	-	10,041	10,254	10,041	10,254
Maltese	r	-	-	5,395	7,574	-	-	5,395	7,574
Maori	r	-	-	13,566	27,965	-	-	13,566	27,965
Mongolian	r	-	-	612	948	-	-	612	948
Nepali	r	4,771	6,074	-	-	-	-	4,771	6,074
	o	-	-	-	-	9,899	9,942	9,899	9,942
Newar / Nepal Bhasa	o	-	-	-	-	4,915	4,921	4,915	4,921

Table 11.2 – continued from previous page

Language	Scr	Scanned Paper		Electronic		Crowdsourced		**All Dicts**	
		Source Words	Total Trans.	Source Words	Total Trans.	Source Words	Total Trans.	Source Words	Total Trans.
Norwegian	r	-	-	-	-	8,050	8,057	8,050	8,057
Panjabi	r	-	-	13,342	24,758	-	-	13,342	24,758
	o	5,641	18,549	15,783	32,538	9,765	10,060	28,224	58,844
Pashto	a	-	-	-	-	382	411	382	411
Polish	r	-	-	56,233	133,036	9,050	9,096	63,007	140,534
Portuguese	r	-	-	790	840	8,457	8,469	9,046	9,204
Romanian	r	-	-	48,088	130,458	8,560	8,597	54,305	137,470
Russian	c	-	-	89,785	213,073	870	874	90,655	213,947
Serbian	r	-	-	55,349	122,710	10,022	10,077	65,166	132,638
Serbo-Croatian	r	-	-	-	-	9,438	9,465	9,438	9,465
Sindhi	a	-	-	-	-	331	345	331	345
Slovak	r	-	-	48,683	122,015	9,315	9,328	55,477	129,485
Slovenian	r	-	-	-	-	9,111	9,115	9,111	9,115
Somali	r	-	-	227	230	6,912	6,946	7,055	7,116
Spanish	r	-	-	73,589	204,754	8,647	8,690	77,465	209,810
Sundanese	r	-	-	-	-	4,537	4,559	4,537	4,559
Swahili	r	-	-	-	-	7,328	7,357	7,328	7,357
Swedish	r	-	-	58,012	116,784	8,098	8,122	62,593	122,770
Tagalog	r	78,081	155,170	24,498	62,505	7,854	7,964	96,476	213,419
Tamil	o	58,236	161,754	-	-	10,119	10,212	66,242	171,228
Tatar	r	1,642	1,862	-	-	-	-	1,642	1,862
	c	-	-	3,795	4,818	-	-	3,795	4,818
Telugu	o	-	-	-	-	10,064	10,177	10,064	10,177
Thai	o	-	-	6,517	14,170	4,594	4,713	10,640	18,732
Tibetan	r	-	-	25,145	55,520	-	-	25,145	55,520
Tigrinya	r	-	-	55	56	-	-	55	56
Turkish	r	114,826	231,795	397,863	572,847	9,073	9,602	493,571	776,089
Turkmen	r	1,824	1,978	68,050	76,282	-	-	69,007	77,528
Uighur	r	1,880	2,131	3,827	4,774	-	-	5,480	6,723
Ukrainian	r	-	-	8,324	14,056	-	-	8,324	14,056
	c	-	-	-	-	10,005	10,013	10,005	10,013
Urdu	r	-	-	15,208	33,832	-	-	15,208	33,832
	a	-	-	-	-	10	10	10	10
Uzbek	r	1,964	2,222	124,280	154,991	4,342	4,481	128,420	160,177
Vietnamese	r	-	-	-	-	4,117	4,270	4,117	4,270
Waray-Waray	r	-	-	-	-	5,033	5,282	5,033	5,282
Welsh	r	-	-	14,586	25,832	7,166	7,176	19,124	31,259
West Frisian	r	-	-	-	-	4,411	4,437	4,411	4,437
Wolof	r	-	-	-	-	36	36	36	36
Yoruba	r	-	-	-	-	1,810	1,855	1,810	1,855

Table 11.2: Data statistics for all dictionaries. The number of unique source sides in each dictionary is given along with the total number of translations. The second column indicates whether the primary script used in the dictionary is Roman (r), Cyrillic (c), Arabic (a), or other (o).

Language	Web Crawls		Wikipedia	
	Total Words	High Precision Words	Words	Pages
Afrikaans	2,204,061	2,074,212	7,296,726	18,904
Albanian	9,127,415	8,958,410	6,388,669	19,860
Alemannic German*	0	0	4,447,599	9,759
Amharic	619,349	562,065	414,578	3,073
Ao*	19,701	-	0	0
Arabic	18,919,351	15,434,228	38,805,091	115,624
Aragonese	2,676,458	2,339,692	2,685,326	15,078
Armenian	3,161,317	3,049,295	5,912,883	30,046
Assamese*	413	0	828,350	1,061
Asturian	329,080	119,837	2,177,174	8,616
Azeri	3,842,179	3,769,223	6,747,026	26,896
Bashkir*	7,425,473	0	818,838	5,112
Basque	7,135,032	7,115,461	24,033,895	93,890
Belarusian	10,570,854	9,632,568	9,726,774	41,657
Bengali	8,295,164	8,159,086	4,998,454	18,603
Bihari*	0	0	68,093	2,250
Bishnupriya Manipuri*	954,790	0	1,394,021	14,986
Bosnian	8,647,129	2,627,513	7,515,961	19,981
Breton	0	0	7,962,606	30,732
Buginese*	0	0	264,962	13,152
Bulgarian	34,042,882	32,684,588	33,926,577	88,436
Burmese*	4,190,851	4,111,818	1,660,176	3,639
Catalan	4,067,627	3,956,588	81,185,339	182,412
Cebuano	1,886,463	1,180,374	2,755,209	52,026
Central Bicolano*	0	0	343,513	4,330
Chinese	2,067,024	59,997	49,808,610	288,528
Chuvash*	0	0	1,259,044	6,534
Croatian*	4,509,460	4,476,599	26,242,455	58,328
Czech	1,553,645	1,539,509	61,572,889	136,353
Danish	6,807,398	6,652,405	13,908,104	40,832
Dari*	3,607,320	-	0	0
Dhivehi*	2,544,804	2,351,921	383,263	1,967
Dutch	24,186,602	23,784,141	89,235,296	299,329
Esperanto*	1,216,714	1,081,055	17,182,282	78,997
Estonian*	5,765,919	5,605,449	14,626,071	50,777
Faroese*	678,443	483,512	925,705	4,691
Farsi	703,507,414	699,678,151	34,957,979	145,609
Finnish	5,607,541	5,132,451	43,164,766	166,371
French	131,582,433	111,885,989	340,158,674	575,923
Galician	1,511,284	1,349,748	24,948,863	52,645
Georgian	3,889,908	3,771,502	8,762,070	42,447
German	58,381,197	51,519,965	314,275,796	488,360
Gilbertese*	311,488	-	0	0
Goan Konkani*	2,360,130	-	0	0
Greek	13,523,704	13,326,526	29,680,446	52,724
Gujarati	1,084,719	1,035,840	3,958,031	3,909
Haitian	49,855	49,845	1,055,107	28,247
Hebrew	10,917,090	10,840,597	52,719,531	83,317
Hindi	31,123,091	30,202,526	16,198,183	25,078
Hungarian	542,736	523,814	69,695,400	127,406
Icelandic	1,161,186	1,156,491	4,457,412	17,469
Ido*	0	0	3,364,125	14,509
Ilokano	0	0	658,476	4,714
Indonesian	5,067,534	4,797,560	26,769,690	83,274
Irish	1,594,775	1,581,142	2,849,017	16,924
Italian	16,875,295	14,134,178	212,715,388	452,758
Japanese	79,015	51,278	117,633,625	296,243
Javanese	0	0	3,469,927	14,105
Kalaallisut (West Greenlandic)*	7,936	0	68,255	1,386
Kannada	1,036,132	1,001,895	8,248,416	6,134
Kapampangan	0	0	520,096	5,233
Kashmiri*	0	0	3,077	84
Kazakh	3,213,297	0	9,001,990	71,874
Khasi*	2,538,370	-	0	0
Khmer*	16,453,655	16,212,496	813,036	1,638

Table 11.3 – continued from previous page

Language	Web Crawls		Wikipedia	
	Total Words	High Precision Words	Words	Pages
Kinyarwanda*	0	0	88,369	1,263
Korean	5,589,281	5,571,812	37,776,582	132,629
Kurdish	4,892,227	392,812	1,348,330	6,351
Kyrgyz	6,335,216	0	974,905	3,237
Lao*	3,674,166	3,674,166	98,824	826
Latin	0	0	9,546,651	68,134
Latvian	36,156,391	35,663,711	9,432,914	33,024
Lithuanian	2,854,697	2,783,953	18,865,990	68,942
Lombard*	0	0	2,717,835	20,604
Low Saxon	0	0	4,700,093	13,106
Luganda*	3,019,265	0	4,844	112
Luxembourgish	1,650,310	991,132	5,132,551	21,735
Macedonian	2,084,421	2,047,306	15,443,536	39,669
Malagasy	0	0	8,089,089	34,431
Malayalam	4,056,931	3,998,017	5,080,980	17,009
Malaysian	1,057,879	751,874	14,064,735	112,780
Marathi	28,215,299	0	2,453,664	21,931
Maltese	1,114,480	1,076,440	1,255,312	1,857
Maori	0	0	144,200	2,188
Mizo*	11,189,501	-	0	0
Mongolian	1,188,640	0	1,991,844	8,211
Montenegrin*	909,499	-	0	0
Neapolitan*	0	0	599,295	6,654
Nepali	3,489,101	0	1,878,168	5,854
Newar / Nepal Bhasa	0	0	810,380	12,505
Norwegian	10,575,063	0	26,372,263	84,347
Norwegian (nynorsk)*	0	0	12,720,136	52,542
Occitan*	0	0	11,401,942	58,717
Oriya*	109,020	0	411,392	2,344
Pangasinan*	0	0	36,697	832
Panjabi	1,955,959	1,919,939	1,019,519	5,382
Papiamento*	1,409,987	1,270,770	167,483	757
Pashto	6,820,583	0	936,001	2,281
Piedmontese*	0	0	3,075,850	40,796
Polish	10,433,634	9,891,731	136,151,261	471,136
Portuguese	21,206,449	19,116,752	129,161,465	404,826
Quechua*	0	0	1,040,087	9,492
Romani*	3,902,611	109,324	39,965	445
Romanian	17,608,197	16,090,347	34,672,327	135,874
Romansh*	3,479,519	3,168,779	511,959	2,575
Russian	1,555,264,838	1,386,337,881	210,652,169	400,797
Sami*	383,959	239,705	172,007	3,692
Sanskrit*	4,617,096	0	768,072	4,534
Serbian	15,194,828	8,654,008	37,575,834	131,854
Serbo-Croatian	0	0	22,695,385	57170
Shona*	2,933,420	2,828,187	92,205	1,121
Sicilian*	0	0	1,501,971	15,454
Sindhi	1,365,512	0	48,095	134
Sinhalese*	1,472,454	1,363,578	1,814,948	2775
Slovak	113,163,058	42,560,917	23,477,764	107,958
Slovenian	2,735,189	2,712,060	20,259,361	57,218
Somali	3,250,014	1,937,974	267,383	1,470
Spanish	913,465,084	861,256,956	232,437,776	374,651
Sundanese	0	0	1,286,508	6,810
Swahili	4,276,643	3,523,452	2,714,133	16,362
Swedish	11,307,825	11,199,464	70,923,386	274,152
Tagalog	3,966,447	3,229,425	434,511	7,298
Tajik*	4,012,768	0	885,687	12,851
Tamil	3,928,554	3,852,001	9,154,660	23,468
Tatar	12,082,366	407	1,535,849	8,410
Telugu	3,254,373	3,220,910	8,769,259	8,841
Tetum*	66,742	24,598	54,292	385
Thai	219,210	219,012	7,431,040	48,679
Tibetan	6,374,651	6,362,553	643,850	2,516

Table 11.3 – continued from previous page

Language	Web Crawls		Wikipedia	
	Total Words	High Precision Words	Words	Pages
Tigre*	5,350,042	-	0	0
Tigrinya	0	0	5,621	106
Tongan*	3,033,770	1,934,709	58,127	405
Tok Pisin*	2,701,541	0	35,249	968
Turkish	14,409,942	13,432,854	30,385,844	89,577
Turkmen	0	0	265,073	1,425
Uighur	1,200,333	0	325,025	1,938
Ukrainian	21,836,916	20,383,329	72,135,536	208,915
Urdu	286,461,259	284,998,217	3,266,533	15,347
Uzbek	8,304,074	1,087,171	5,368,879	71,081
Valencian*	599,617	-	0	0
Vietnamese	2,468,179	2,466,133	53,471,136	194,374
Volapuk*	0	0	15,308,318	97,588
Walloon*	0	0	627,486	3,123
Waray-Waray	0	0	2,858,127	102,823
Welsh	6,573,628	6,565,494	4,414,153	28,066
Western Panjabi*	0	0	1,598,223	19,589
West Frisian	1,766,944	0	5,014,518	13,978
Wolof	0	0	230,337	943
Yoruba	0	0	473,264	23,006
Zazaki*	0	0	255,262	2,380
Zulu*	530,705	427,748	23,095	433

Table 11.3: Amount of monolingual online newspaper crawls and Wikipedia data, by language, after tokenization and pairing with English comparable corpus. The high precision web crawl data contains only text identified by an automatic language identification system as a particular language. High precision dataset statistics are omitted for languages for which no language ID system was available. Languages marked with a * indicate those for which we do not have a bilingual dictionary.

Chapter 12

WADE Analysis: Comparison of the use of Automatic and Manual Word Alignments

For our WADE analyses, incorrect word alignments over our *test* sets are also problematic because the analysis is based entirely upon them. That is, if a word, f_i, in a sentence in our test set is incorrectly aligned with a word, e_j, in the corresponding reference sentence, WADE will incorrectly indicate an error in the output machine translation if the decoder does not produce word e_j from f_i. In order to estimate the effect of incorrect test-reference alignments on our WADE analyses, we manually align a subset of our test sets.[1] We then compare the result of a WADE analysis using manual alignments versus alignments estimated using a high resource and low resource automatic alignment model.

We use MTurk to gather manual word alignment annotations for a subset of our Spanish and Hindi test sets. We began by manually aligning 10 Spanish-English and Hindi-English sentence pairs ourselves and used this small set of gold alignments as a means to identify qualified MTurk workers who provide high quality annotations. Two local native Hindi speakers each annotated ten Hindi-English sentence pairs. Following previous work (Graça *et al.* (2008); Kruijff-Korbayová *et al.* (2006); Melamed (1998), among others), we measure agreement as follows, where A_1 are the alignments given by the first annotator and A_2 are the alignments given by the second annotator:

$$Agreement = \frac{2 \cdot |A_1 \cap A_2|}{|A_1| + |A_2|}$$

Our two Hindi annotators achieved 91.5% average agreement, and we used the union of their alignments as a gold standard. For Spanish, we used a single set of gold standard alignment annotations.

Because manual word alignment is a difficult task (Melamed, 1998), we only annotate sentence pairs for which both the source and target are no more than 20 words long. Following Och and Ney (2003), we allow workers to indicate both 'sure' and 'possible' alignments. Sure alignments indicate a direct correspondence and possible alignments ambiguous or loose correspondences. We provide workers with initial sure and possible word alignments based on intersection alignments and the grow-diag-final heuristic, respectively, and ask them to correct the alignment links. To generate the initial alignments, we concatenate full training sets with our test sets and run GIZA++ in both directions. For the purposes of evaluating the quality of a worker's alignments compared to our gold set, we measure alignment precision, recall, and F-measure using the union of annotators' alignments. We measure each as follows, where G is the set of gold alignments and A is the set of workers' annotated alignments:

$$recall = \frac{A \cap G}{|G|} \quad precision = \frac{A \cap G}{|A|} \quad FMeasure = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

In comparison with the small set of Spanish-English gold word-aligned sentence pairs, the average precision, recall, and F-measure of the initial (GIZA++ and grow-diag-final) alignment set are 79.5%, 81.8%, and 80.3%, respectively. For Hindi, the average precision, recall, and F-measure are 74.4%, 78.8%, and 76.0%, respectively. We give these initial alignments to workers to correct. We expect that good annotators will have a higher agreement with our gold annotations than the automatic alignments do.

For both Spanish and Hindi, we post word alignment HITs for the ten sentence pairs with gold standard annotations and ask up to 25 workers to complete them. For Hindi, the mean F-measure of the 250 annotations was 0.81 with a standard deviation of 0.17. For Spanish, the mean F-measure of the 250 annotations was 0.76 with a standard deviation of 0.19. We identified those 12 Hindi-English and 6 Spanish-English workers who annotated at least five sentence pairs and achieved an average F-Measure score of at least 0.7 and approved those workers to do additional word alignment tasks. Our final set of annotations contains manual alignments for 500 Hindi-English sentence pairs and 500 Spanish-English sentence pairs. We paid workers \$0.35 per alignment task and had each sentence pair annotated by two workers. The total cost of annotation, including the HITs used only to identify reliable workers, was \$962.50. We paid an effective hourly wage of \$8.34 and \$6.77 and workers spent, on average, 3 minutes and 2.5 minutes for the Hindi and Spanish HITs, respectively. Across all 500 Hindi-English HITs, the average agreement (defined above) between each pair of manual alignments is 86.2%, which is only slightly less than the average agreement, 91.5%, between our two expert, local annotators over the set of ten gold aligned sentences. For the Spanish-English HITs, the average agreement is 87.5%.

We rerun several WADE analyses using the subset of test data with manual alignments and compare the following:

1. Test-Reference word alignments using a low resource (LR) alignment model estimated over $1,000$ sentence pairs
2. Test-Reference word alignments using a high resource (HR) alignment model estimated using our full training sets for each language.
3. Manually annotated test-reference word alignments gathered on MTurk.

[1] We also use these manual word alignments in Chapter 7.

	Correct	Freebie	Seen	Sense	Score	Total Number of Alignments
Hindi						
Automatic (LR) Test Set Alignments	31.4	0.4	29.7	30.6	7.9	5,450
Automatic (HR) Test Set Alignments	32.5	0.4	30.4	28.2	8.5	5,372
Manual Test Set Alignments	31.2	0.4	28.0	31.2	9.3	5,827
Spanish						
Automatic (LR) Test Set Alignments	34.9	3.7	27.7	24.9	8.7	8,001
Automatic (HR) Test Set Alignments	37.7	4.1	23.7	23.5	11.1	7,492
Manual Test Set Alignments	33.8	3.7	20.7	31.7	10.0	8,419

Table 12.1: A comparison of WADE results over a subset of the test sets for which we have both automatic and manual word alignments. For both Spanish and Hindi, the analysis is done over 500 sentence pairs.

Table 12.1 shows the results. First, we note that the WADE Hindi-English analysis based on low resource does not vary by much in comparison with that based on high resource test-reference word alignments. Similarly, **there is little difference between the analysis based on manual word alignments and those based on automatic alignments**. Although the percent of automatic alignment links that are marked as correct by WADE when using an automatic word alignment between the test and reference sets is higher than when considering manually annotated alignment links, there are about 8% more manual alignment links than automatic. Thus, the raw number of alignments which are annotated by WADE as correct is actually higher under the set of manual links. The main observation, however, is that the general trends gleaned from each set of analyses is consistent: for Hindi-English translation, most of the mistakes are of type SEEN or SENSE, and many fewer SCORE errors are made. This is reassuring and indicates that, even if we do not have enough parallel data from which to estimate a high quality automatic word alignment model, a WADE analysis based on noisy word alignments yields a reliable perspective on the types of errors that exist in a set of machine translations.

The discrepancies between the WADE analyses using different alignments are larger for Spanish-English than Hindi-English. However, the major trends are consistent, whether we rely upon manual or automatic alignments: SEEN and SENSE errors are a big source of error, and there are fewer SCORE errors. When moving from lower quality (low resource automatic) to higher quality (manual) alignments, the number of alignments classified as SEEN errors decreases, as expected. At the same time, the number of alignments classified as SENSE errors increases as many previous SEEN errors become SENSE errors under a more accurate set of word alignments. As we found in the Hindi experiments, there are more manual word alignment links than automatic.

Chapter 13

Bilingual Lexicon Induction

13.1 Contextual Vector Projection Dictionaries

Throughout the experiments in Chapter 4, we assumed access to some bilingual dictionary, which we used to project source language contextual vectors into the space of target language contextual vectors before scoring contextual similarity. In this section, we compare using a *probabilistic* dictionary for projection with using a *non-probabilistic* dictionary.

One of the data scenarios which we are most interested in is the case where we have access to a modest amount of parallel training data, and we wish to use monolingual resources to improve our statistical MT model. In this situation, we can use automatic word alignments to extract a dictionary. Along the way, we can also collect counts of how many times each source and target word pair are aligned and use relative frequencies to estimate conditional probabilities, $p(trg|src)$. We can then probabilistically project each context vector. For example, say we have the following context vector for some Spanish word: [*blanco*: 2, *casa*: 1, *bueno*: 2]. Our probabilistic dictionary may tell us that *blanco* translates as English *white* with probability 1.0, but *casa* translates as *home* with probability 0.25 and *house* with probability 0.75 and, similarly, *bueno* translates as *good* with probability 0.6 and as *ok* with probability 0.4. Our probabilistically projected context vector would then be: [*white*: 2, *home*: 0.25, *house*: 0.75, *good*: 1.2, *ok*: 0.8]. Using a non-probabilistic dictionary, the resulting context vector would be: [*white*: 2, *home*: 1, *house*: 1, *good*: 2, *ok*: 2]. Intuitively, using a probabilistic dictionary for projection should allow us to give less credit to infrequent translations and also to avoid projecting too much mass for source words with many target translations.

For this set of experiments, we use the Indian languages corpora described in Post *et al.* (2012). In particular, we use the parallel training sets for Tamil, Bengali, and Hindi to extract probabilistic dictionaries. We compare using intersection and grow-diag-final GIZA++ alignments (Och and Ney, 2003). In all cases, we extract contextual vectors from both crawled data and Wikipedia data and project them separately, using the given projection dictionary. We then train and use a reranker based on *only* the two contextual feature vectors. As a result, top-k accuracies are much lower than when we use more complete feature sets. However, by isolating the contextual similarity features, we may make comparisons across context vector projection dictionaries.

Table 13.1 shows results using several different contextual vector projection dictionaries. In general, it's clear that using both the original Mechanical Turk dictionaries and those derived from the word aligned training data is better than either either alone. Additionally, intersection alignments tend to result in higher top-1 performance than using the less precise but higher recall grow-diag-final alignments. Conversely, the grow-diag-final alignments yield higher top-10 and top-100 performances. Using probabilistic projections tends to slightly hurt performance, at least for this low resource condition.

13.2 Comparison of Temporal Signatures

In Section 4.2.1.2, we described our method for estimating and comparing the temporal signatures of a pair of words. In our bilingual lexicon induction experiments, we found temporal similarity to be a weak but useful signal of translation equivalence.

Here, we compare two methods for measuring the similarity between a pair of temporal signatures. Recall that word w's temporal signature is an M-dimensional vector of counts of the number of times w appears in each of M time periods, and each count is normalized by the sum of all counts. In our web crawl data, we have dates associated with articles and each time period corresponds to a single day. In the experiments presented above, we used a simple dot product to estimate similarity. However, this metric may not be ideal. For example, it may be the case that an event is covered extensively in news stories in one language on a particular day or week but that same event is not picked up in the newspapers of another language until some time after. We have attempted to account for such a temporal shift by using a sliding window of three days when populating temporal signatures, but the sliding window would not be able to smooth over longer shifts, which would result in poorly aligned temporal signatures.

Here, we compare the performance of using temporal signatures based on normalized raw frequency counts with a more sophisticated transformation of the signatures.[1] In order to account for temporal shifts, or imperfectly aligned time series, we compute the discrete Fourier transform (DFT), $\mathcal{F}(X)$, of each M-dimensional series X:

$$\mathcal{F}(X)_k = \sum_{m=0}^{M-1} x_m e^{-i2\pi mk/M}$$

where k is an index over the output transformed vector.[2] The Fourier transform assumes a series of equally spaced samples. In our datasets, however,

[1] In all experiments here, we use the same three-day sliding window that was used in earlier experiments.

[2] Thanks to Mike Carlin for, after all those years, teaching me about discrete Fourier transforms.

Dictionary	Probabilistic	Including MTurk	OOV Accuracy Top-1	Top-10	Top-100
Tamil					
MTurk	No	Yes	5.50	13.61	28.02
Intersection Alignments-Based	No	No	6.71	17.53	34.61
Intersection Alignments-Based	No	Yes	**8.10**	20.64	37.28
Intersection Alignments-Based	Yes	No	3.73	10.60	26.42
Intersection Alignments-Based	Yes	Yes	4.62	12.29	28.44
Grow-diag-final Alignments-Based	No	No	7.28	22.36	40.15
Grow-diag-final Alignments-Based	No	Yes	7.77	**23.01**	**40.67**
Grow-diag-final Alignments-Based	Yes	No	4.90	13.17	30.39
Grow-diag-final Alignments-Based	Yes	Yes	5.22	13.92	30.78
Bengali					
MTurk	No	Yes	7.05	17.45	35.45
Intersection Alignments-Based	No	No	9.36	22.79	41.76
Intersection Alignments-Based	No	Yes	**11.19**	25.61	44.53
Intersection Alignments-Based	Yes	No	6.29	15.44	33.03
Intersection Alignments-Based	Yes	Yes	8.20	18.88	37.23
Grow-diag-final Alignments-Based	No	No	6.61	27.76	49.06
Grow-diag-final Alignments-Based	No	Yes	7.09	**28.73**	**49.92**
Grow-diag-final Alignments-Based	Yes	No	7.49	19.74	39.80
Grow-diag-final Alignments-Based	Yes	Yes	8.97	21.89	41.25
Hindi					
MTurk	No	Yes	7.01	17.80	39.15
Intersection Alignments-Based	No	No	13.44	30.25	54.40
Intersection Alignments-Based	No	Yes	**14.42**	31.51	55.58
Intersection Alignments-Based	Yes	No	7.13	18.60	39.63
Intersection Alignments-Based	Yes	Yes	8.47	20.95	43.06
Grow-diag-final Alignments-Based	No	No	6.96	23.85	54.47
Grow-diag-final Alignments-Based	No	Yes	7.08	**24.58**	**54.85**
Grow-diag-final Alignments-Based	Yes	No	8.24	21.45	44.87
Grow-diag-final Alignments-Based	Yes	Yes	8.95	23.37	46.91

Table 13.1: Comparison of dictionaries used to project context vectors. Evaluation is over all source language words in the MT development set. Gold standard translations are taken from the automatically word aligned development set and its reference translations as well as all of our available bilingual dictionaries for each language. If *any* gold standard English translation is found in the top-k ranking for a given source word, we consider that source word to be accurate in the top-k.

we frequently don't have data for every day in the range of a given time series, particularly for lower resourced languages. In the experiments here, we simply ignore days with missing data. However, we experiment with low as well as high resource languages, for which there are very few missing samples.

The Fourier transform is an equivalent representation of our original time series as a sum of complex sinusoids. Such a representation is common in signal processing for characterizing slow versus fast modulations in the original time series. In particular, the *magnitude* of the DFT, $\hat{X}_k| := \mathcal{F}(X_k)|$, quantifies the strength of these modulations regardless of any time shifts in the original time series. It is this temporal phase-invariance of the DFT magnitude that makes it useful for comparing pairs of time series that may have similar overall temporal structure but may be shifted versions of one another. Therefore, to compare pairs of temporal signatures, we use a the dot product of the DFT magnitudes, i.e.:

$$sim_{temp_{dft}} = \frac{\hat{F} \cdot \hat{E}}{||\hat{F}||||\hat{E}||}$$

where $\hat{F} = |\mathcal{F}(F)|_1^{\frac{k-1}{2}}$ and $\hat{E} = |\mathcal{F}(E)|_1^{\frac{k-1}{2}}$, ignoring the DC terms.

We experiment with three language pairs, Spanish-English, Indonesian-English, and Cebuano-English, to test the effect of using Fourier transforms in measuring temporal similarity. Spanish-English is a high-resource language pair, and we have nearly a billion words of Spanish time-stamped web crawls. Indonesian-English is a medium-resource language pair; we have about five million words of web crawled data. In contrast, we have only about one million words of Cebuano web crawled data.

In order to compare the effectiveness of using standard temporal signatures versus their Fourier transforms, we experiment only with words for which we have reasonably strong temporal signatures. That is, our estimates of the temporal signatures of words that appear in our comparable corpora infrequently are sparse and unlikely to be meaningful whether we use Fourier transforms or not. Therefore, for each language pair, we collect

Signatures	Source Test Words	Target Candidate Translations	Mean Rank	Mean Rank Percentile	Accuracy Top-10	Accuracy Top-100
Cebuano						
Norm Frequencies	202	4,452	2,055	46.2%	0%	2.5%
DFT of Norm Frequencies			3,148	70.7%	0%	1.0%
Indonesian						
Norm Frequencies	741	18,376	6,608	36.0%	0%	0.3%
DFT of Norm Frequencies			11,225	61.1%	0%	0%
Spanish						
Norm Frequencies	1,000	20,050	5,272	26.3%	1%	3.4%
DFT of Norm Frequencies			7,583	37.8%	0%	0%

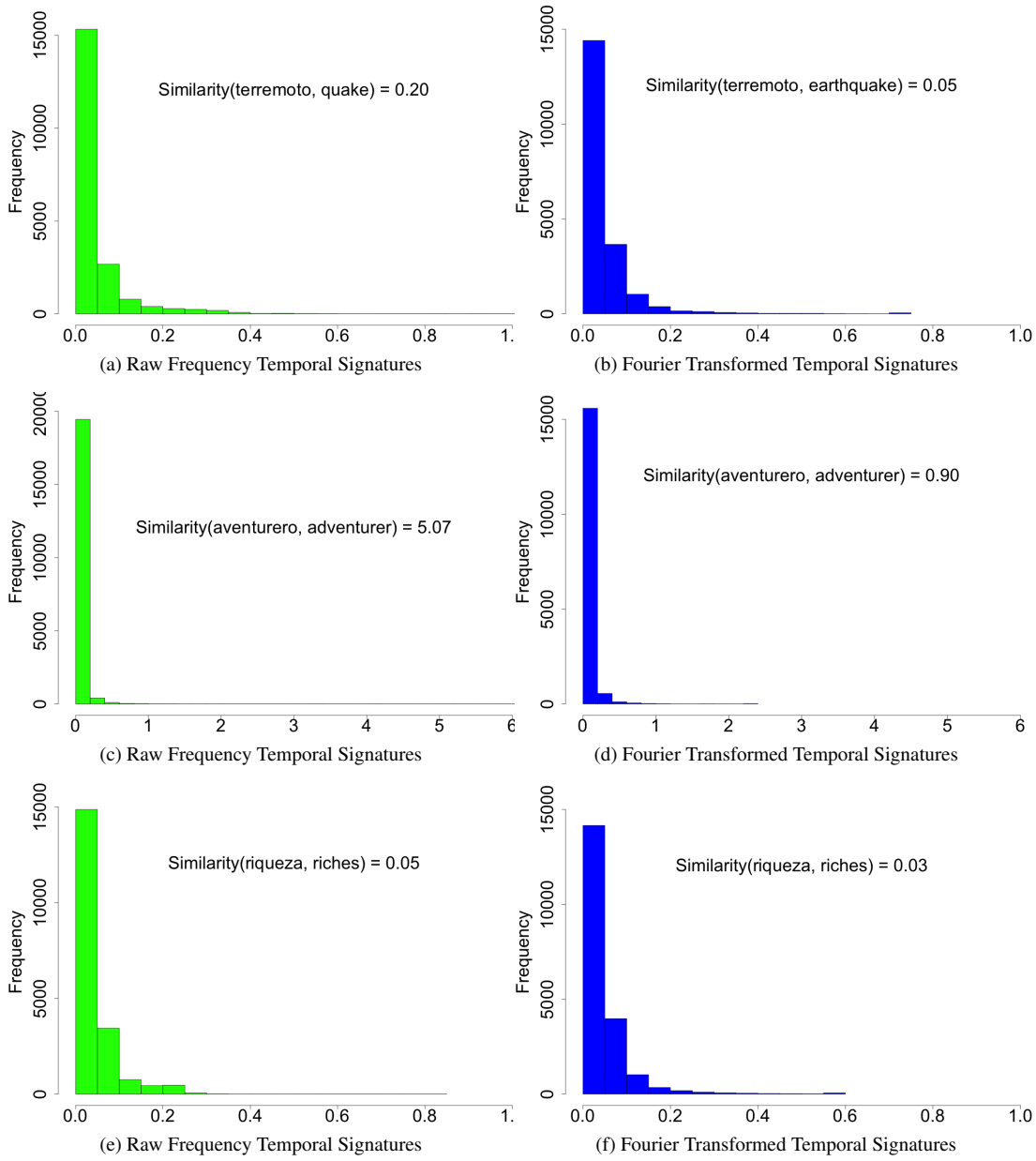
Table 13.2: Translating ranking performance using temporal signatures comprised of normalized frequency counts versus their Discrete Fourier Transforms.

temporal signatures for only those source and target language words that have a monolingual frequency of at least 300 in our comparable corpora. We then prune the source language words to those which, given our bilingual dictionaries, have a known translation in the set of frequent English words. This results in a test set of 26,349 Spanish, 741 Indonesian, and 202 Cebuano words. We further prune the Spanish set by randomly sampling 1,000 words. For both raw and Fourier transformed temporal signatures, we estimate the similarity between a pair of temporal signatures using dot product. We compare each source language word with each English candidate translation. There are about 20, 18, and 4 thousand candidate English translations in the search space for our Spanish, Indonesian, and Cebuano experiments, respectively. We rank these English candidate translations by their similarity scores with test source word and evaluate using mean rank and top-k accuracy.

Before considering aggregate results over each set of test source language words, we spot check the impact of using Fourier transforms on several example word pairs. Figure 13.1 show histograms of the similarities between a given source language word and the English candidate translations. The similarity between the source word and the highest ranked correct English translation is given on each subfigure. Note that we normalize similarity scores by the sum of all scores and plot the scores times 1,000 for readability. We sample three Spanish words of varying burstiness: *terremoto* (*earthquake*), which is has a high burstiness, *aventurero* (*adventurer*), which has an average burstiness, and *riqueza* (riches), which has a low burstiness. The pairs of similarity histograms, in general, look similar. The Spanish words *terremoto* and *riqueza* have very low similarity scores with their correct English translations, *quake* or *earthquake* and *riches*. However, the Spanish word *aventurero* has very high similarity, 5.07, with its English translation *adventurer* using raw frequency-based temporal signatures. The similarity score using Fourier transformed vectors is 0.90. Although this value is high, it is not as discriminating.

Figure 13.2 provides an additional illustration for comparing pairs of raw frequency based temporal signatures and their Fourier transforms. The top two figures, (a) and (b), show the raw frequency and Fourier transformed signatures, respectively, of *terremoto* paired with its correct translation, *earthquake*. Both pairs of signatures show similarities. Note that the lower dimensions of the Fourier transformed signatures have higher magnitudes than the higher dimensions, indicating the presence of stronger, slower modulations over time. The bottom two figures, (c) and (d), show the raw frequency and Fourier transformed signatures, respectively, of the same word paired with an incorrect, though, in some contexts, topically related, English candidate translation, *strength*. Although both pairs of signature show less similarity than was present for the correct English translation, the raw frequency signature shows even less than the pairs of Fourier transformed signatures, which is the desired behavior.

Table 13.2 shows aggregate results across our entire test sets of our experiments measuring the impact of computing Fourier transforms of our frequency-based signatures before measuring temporal similarity. For our task and datasets, using Fourier transforms hurts performance slightly. For all language pairs, correct English translations are ranked higher in the set of all ranked candidates when we use raw temporal signatures than when we use their Fourier transforms. As expected, performance is higher for Spanish-English than the other language pairs. Because our web crawls a much larger for that language pair than the others, the temporal signatures that we estimate are more complete and meaningful. However, given the general unpromising results presented here, we do not use Fourier transforms of our temporal vectors in place of raw frequency-based vectors as a feature in our bilingual lexicon induction models.

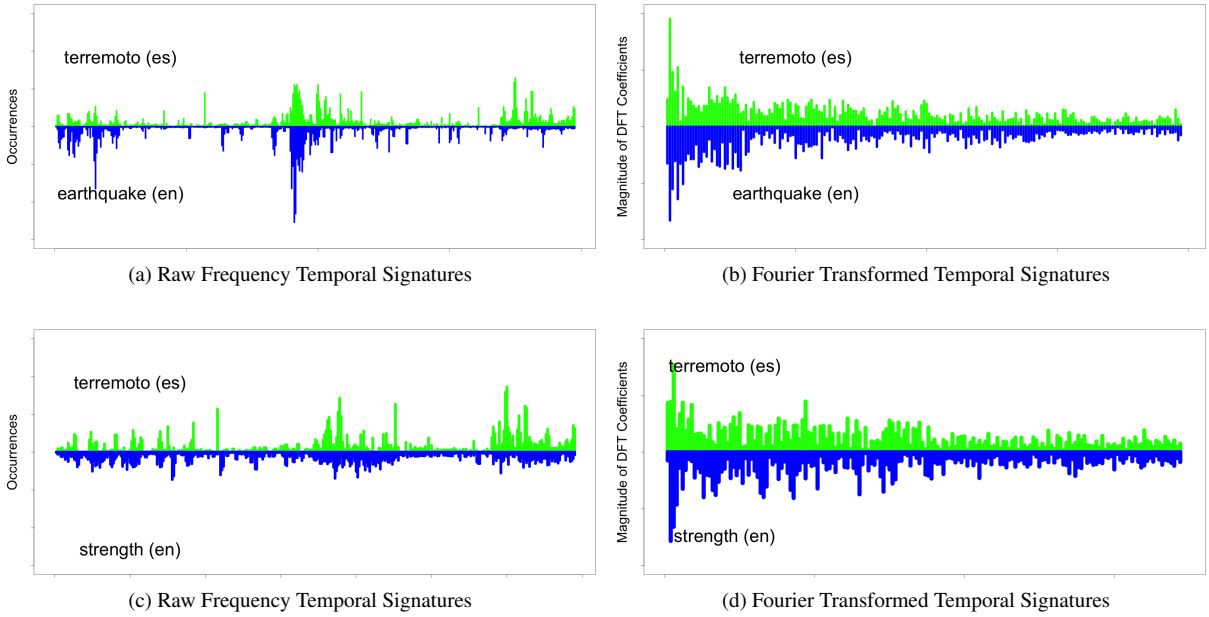


(a) Raw Frequency Temporal Signatures (b) Fourier Transformed Temporal Signatures

(c) Raw Frequency Temporal Signatures (d) Fourier Transformed Temporal Signatures

(e) Raw Frequency Temporal Signatures (f) Fourier Transformed Temporal Signatures

Figure 13.1: Histograms of similarity scores between several Spanish test set words and all English candidate translations. Similarity scores are shown on the x-axis and the frequency of different scores is shown on the y-axis. Frequency-based temporal signatures are used to compute similarity scores in the histograms on the left (green) and their Fourier transforms are used to compute similarity scores in the histograms on the right (blue). The Spanish word *terremoto*, which means *earthquake* or *quake*, is the source language test word in (a) and (b). In (c) and (d), it is *riqueza*, which means *riches* or *wealth*, and in (e) and (f) it is *alcanzaron*, which means *reached*. On each plot, we give the similarity between the source language test word and the highest scoring correct English translation.



(a) Raw Frequency Temporal Signatures

(b) Fourier Transformed Temporal Signatures

(c) Raw Frequency Temporal Signatures

(d) Fourier Transformed Temporal Signatures

Figure 13.2: Comparison of temporal signatures for the Spanish word *terremoto*. Both raw and Fourier transformed signatures are compared with both a correct, *earthquake*, and incorrect, *strength*, English candidate translation.

Chapter 14

Zero-Parallel Data Translations

Table 14.1 gives translations of the first 4-5 sentences of foreign language Wikipedia pages on *Barack Obama*. A blue word indicates a dictionary translation, red indicates a transliteration, and green indicates an induced translation. Black words are induced translations that turn out to have exactly the same string as the input. Orange indicates that a word was both the best transliteration and in the top-10 induced translations.

Language	Translation Output
Nepali	senator barack obama (birth : 4 august , 1961 america) ryan washington is — he this country first folio (african americans) washington is . he 20 january , 2009 for day position newly shy take is done . senator barack personnel herald senate and 2008 in america blamed for viewing parties democratic candidate was . senator barack l aung san in 1991 graduate formed , where he barack obama l slaves first african president also was . 1997 from presidential senator senate transcript three complete to do before communal senator obama in the form of work done and citizens authority responded in the form of spherical did . 1992 from presidential upto he harvard law university constitutional law amendment work also did san . 2000 in american house af representatives vargas get for doing unsuccessful in 2003 after january his sight born obama was . away seats in primary victory achieved in 2003 and did not directly for elected was . 109 n trinidad democratic blank oval in the form of traditional he hat-htiarmathi control and federal treasury in use more public than doing support plans in making coperation . did he eastern europe , middle east and africa royal journey also . went 110 n trinidad and tobago change change , climate change , climate change and tobago defeat nepalese closely related welfare vidheyk-nirmana-ma he did coperation . world war important contribution did want delivered year graduate 2009 for nobel prize exchange
Somali	barack obama it is a man of juvenile and you to cargo many mediation between and usa indiana it is 47 . old the he was born hawaii in the region last month 27 juli seeds , year 1961 dii. the birth just father people black is from came from nairobi of kenya and mother of white people is from , watch region areas of country america him and ladies barack obama , the have the two girls and with with called mail and sasha the father the from he left him and two years of old to the he studied from give leads the osaka harvard . adding the the after he returned to the land from came from of kenya . illinois after , senator obama the mother is she married a man from came from the country indonesia. and then they to have moved from srilanka are mostly migrants and is in nominee while to he stayed out-it in the one 1971dii until 1969 . obama , when the after to , he returned hawaii region of the mother in the end it was , the there with avoiding lived in mother he was born of was white .
Uzbek	senator barack obama ii (pronunciation nava : barak obama ; a second time 1961-yili 4-avgust) the united states 44- as well as of the present presidents . buganda chicago maine task list senator bavaria . 2009-yil 9-oktabrida totalitarian policy for nobel prize by thailand . all open barack obama says president barack obama , white it
Azeri	barak huseyn barack (. ing barack obama ii keyes ; 4 august 1961) - congress the united presidency 44th president . 2009-cu in nobel peace award said . vital 4 august 1961-ci in hawaii hilo born the barak huseyn barack obama father also ordinary huseyn barack obama (father) , cowboys his mother ordinary the puck carey obama imam . obama the his father with obama the the mother hawaii university am were . barak barack 1983-c in colombia radcliffe he graduated from the and 1985-ci in obama salinas in that place obama and oldenburg obama obama living conditions improve for hasan the caucasian koller started to work . 1991-ci in barack harvard law school nights .

Language	Translation Output
Tamil	guo barack obama (bezos barack obama , transcribe : baptismal immerse baptisms , born : august 4 , 1961) , of america 2008 republic leader elections mls democracy party baptisms . now he submersion obama state support younger as members . those who are america allawi africa america race from first republic leader in manjrekar , and sent house fifth abraick-amerike native in manjrekar for them present . colombia immerse harvard leber immerse slam received obama politics world insead in front of obama south in society obama (page opinion) public law obama numb . 1997il obama state basu unctad 2004 time in power he was there . 1992 first 2004 time chicago university leber college ito numb . 2000il america lobbyists elections defeat to get 2003 disclosures obama aacsb obama begins . obama in the state legislative as members will be cipher , 2004l democracy party national obama he did obama national level attention baptisms . after november 2004 from that year america parliment elections 70 % immerse greatness obama joined . 2004il democracy party fers aru were period in power obama bashir . central government weapons control akbar law tupac . 2006il democratic party isga layne later on election fraud , climate bitmap , blix extremism , before immerse orphanages likewise automaker related laws was written by february . 2007il republic leader election competition declare obama iraq war adrs audacity , power freedom delusions , outside aacsb politicians on effect reduced , to everyone hygienic program as own important principles that said .
Albanian	barack obama downturn hussein obama (born on 4 august 1961 in honolulu , hawaii) is president of 44-t to him to united states of america . barack obama is winner of price in behalf of nobel prizes in year 2009 . he is president first afro-american . ¡p¿ in year 2004 he was elected in american satin in illinois . candidature such as president in campaign of elections and declared in 10 in a word 2007 in wheelchair (illinois) . on 3 june 2008 obama reached , in accordance with of data to cnn , derive and necessary to vote for to nominated as candidate for president , within party that his , by left at thereby lugar and mcmillan americans , wife of old to us president bill clinton . in 27 august 2008 barack obama was moria on way out of official to epp and democratic party such as obama for president .
Bengali	barack obama hussain , junior (english language : spoken spoken prize , jr) . (births 4th august , 1961) nobel peace reward winner and united states joyner obama . president he democratic of the party member . of before he united states quang illinois province selected representative or senator responsibility celebrate did . obama 2008 of the year 4th november was held united states joyner mckinney in the election winner was and 2009 of the year 20th january oath did receive . october 9 , 2009 date obama in peace nobel prize give do is . primary life barack obama united states joyner hawaii province capital honolulu birth did . his father in kenia isolates nation barack obama hussain cinri (spoken spoken prize ,) sr. was one economist and his mother ann obama (windus reelected) was united states obama (mainly english o irish) . obama howai-manoa father of university to read during the time an unsatisfied with his identity o marriage is . 2 obama year in age his baba-mayer jolie happen . mother obama after indonesian lolo dunham (javanese : obama) interracial marriage did . obama many of childhood time span indonesia . 10 year in age he his obama nana-nanir near came came . in the next obama harvard university from in law degree gain did
Welsh	barack hussein obama ii is 44ain president the united states . a 'n member of 'r democratic party . the was obama in senator from illinois from 2004 to 2008 . was elect its presidential election in november 2008 against john mccain . the was oprah winfrey and ted kennedy in his support . he is 'r african american first available in order to elect its president in the united states . took in university and columbia university the law , harvard . at after published her that for stand for the presidency in february 2007 , published also its edmund in order to see his soldiers country in pull out of iraq and yan . its history early born obama on 4 august 1961 in honolulu , hawaii ; in son in order to barack obama , high , black man from napa muscat dunham in kenya and ann dunham , americans skin jones of wichita , kansas . met her parents while they in hawaii university .
Bosnian	nobel hage biden i (fon . hound khatib biden i ; born 4. august 1961 in honolulu , hawaii , now -) is american politician and 44 . president united american country , therefore former member american senate (representative country illini) . biden is was candidate in front of democratic parties to elections for president united american country , which were are rune november 4. 2008. years , on which is won republicans john mccain . biography born from marriage keyes father and american mothers , most its own life translated by is in honolulu , to hawaiians . in period since 1967 . - 1971. years lived is from mother and specter in jakarta , indonesia , where is elementary and school . graduated is right at harvard university and columbia .
Latvian	biden hussein obama , youngest (, born) is forty fourth , as well as current u.s. president . he is first peachtree , which is interprets this post . out of 2005. annual by ratzinger presidential nominee obama was illinois state senator . barrack obama inauguration ceremony was . obama is hiatus columbia university and harvard obama school (harvard law school) , a he was first harvard law api beyonce president . when obama was public employee chicago , where he mission as well practice that human rights lawyer . out of 1997. as 2004. the obama on three presidents was elected barack obama . he as well out of 1992. as 2004. the chicago university obama school press pcf constitutionalist right . 2000. year he ballon place u.s. higher vichy , but 2004. year november obama was elected u.s. senate .
Indonesian	onyango hussein obama (ii ;) there was president american union which these days to hold and to form president america the union ke-44 . onyango to become since 20 january 2009 to succeed george walker bush . before he is junior senator from illinois and then to win in election president 2008 to 4 november 2008 . he went home and brought his title with him by 2009 , obama announced such as winner gift from god honor peace prize due to conduct of the relations of one state with another by peaceful means to promote international to to solve masalah-masalah international . obama is descent afrika-amerika the first president to be the american union after previously to form descent afrika-amerika the first which yudhoyono by a political party and definitions which can be used as a principle for the whole processing of the language matter large american to to be president . columbia university graduate and school unwritten law in as it is based on custom harvard university ; in there he to hold such as president harvard law review , obama to work such as coordinator society and to hold such as civil rights lawyer to be appointed before senate chicago as long as three times to start 1997 until 2004 .

Language	Translation Output
Romanian	barack obama second dcp (pronounced in english / churchman obama obama / ; n . 4 august 1961 , honolulu , hawaii , the son of senator barack obama , sr. born in province kenyans obama , with ethnic group luo and al by ann obama born in wichita , kansas) there is al 44-lea chairman al the states united (first afro-american high in this position) , for of the elections held on 4 november 2008 , by the side of envy obama as to vice-president) . was has been investing acting on 20 january 2009 and l-a changed by his predecessor george w. bush . the son of one kenyans , barack hussein obama , sr. and al of a u.s. , ann obama alex , obama was spent a large part of childhood in honolulu , hawaii . in six as far as decade of lived in jakarta , with his mother and father with the object of cruel , indonesia . graduate al graduated columbia and al faculty of right leg from harvard , to a run for the sake of entrance into harta public so as to join a al the senate state illinois between 1997 and 2004 , obama was wrought such as conciliator tce , docent university and such as solicitor specialized in the defense rights leaders .
Serbian	barack hussein obama - (ii) - (, born august 4. 1961 in honolulu , on jiao , in sad) is current american president , chosen on elections year 2008. . member is democratic party . of 2005. to 2008. was is younger senator from illinois . born is in honolulu on jiao , as child molestation and carjacker . grown is on different places . bigger part own childhood spent is on jiao , and four year in youth is spent in indonesia . graduated is political science on columbia university , and the title lustig usurper (- (j.d.) -) acquired is on harvard . before but which is became senator in senate states illinois (between 1997. and year 2004.) , dealt with is social work and grillo , in areas civil rights . their campaign for entry in american senate is started 2003. year . on elections is won republican alan , see also gatsby with great majority of 70 % votes . in process presidential campaign , obama is promised , that will in case his win on elections seen to significant changes in washington and that will to him one of priorities to be retreat american troops from iraq .
Turkish	barack obama def (, birth . 4 august 1961) the united states of 'nin state bond presidency . 4 november 2008'de being made the us 2008 presidential elections 'nde major 'nin 44. president elect and 20 january 2009 history such acts as a george w. bush 'tan obrador . the us history the multiracial state bond presidency . life the us 'nin hawaii rochelle in the city of mauna obama medicine center of gravity 'nde the world was . with self as much as name carrying father kenya 'nn nairobi region 'ndeki ruhr naga obama residential district in was born and brought up in the lap of luxury any obama . the refrain from obama of muslim when (in private muslim) sexist all the next world and valdes to cause has been . on the other hand the view muslims obama or obama fairness to cause has been his mother ann yeats in that case kansas state wiccans in the city of was born and brought up in the lap of luxury any olmos . obama 'nn old lady and father , his father's foreign student in that hawaii 'de met and got married . new married obama binary 2 years old divorce . father boston 'a step by step cambridge university 'nde doctorate degree to make and 1965 in kenya 'ya back . returned to his mother in that case once again any foreign student one cadres obama sanya 'yla second any marriage to make .
Ukrainian	baranov obama k. in obata ma (; writer . 4 august 1961) & the ; - american poet , 44-i president usa , former senator from illinois state in congress usa . first was elected to senate state illinois in 1996 . in november 2004 with when bmx barack obama hinojosa political opponents on elections - and was reelected congress united states america , after assassination attempt obama in parliament representatives usa in 2000 . in democratic party got nomination on incumbent president and 23 august 2008 r. announced the choice candidacies joe biden on obama vice-presidenta . for results elections 4 november 2008 gained r. victory above augment candidate jon obama and 20 january year 2009 was first in history country president tryout usa . biography barack obama born in honolulu , hawaii state , in his barack hussein obamy-starshogo that anna oremans . joho father origins with settlement allegory province nyanza republic kenya , and have with place wichita american state kansas . father chaplin under hour training in cochlea university , where obama-starshi later as dozens student . when obama two years ago joho father at first mullis , and afterwards officially lakota divorce . joho father began training in pedagogical in harvard university , where he gained blasts degree doctor that returned to kenya . have obama married marry yanukovych , this time for indonesian students inhabiting acu , from which in being born daughter .
Hindi	barack hussein obama (birth : august , 1961) america of chatrier president of a nation is . they the country its first saharan (african american) president of a nation is . they 20 january , 2009 ko president of a nation post of oath lee . obama biden region from smallest gaylord and 2008 among america of president of a nation post of for godot morcha its candidate was . obama harvard lo school from 1991 among graduate became , where they harvard lo voight of before african american adarsh also stay . 1997 from ussr biden normand among three complete zemin does its east obama ne aquatic organiser of shape among work did is and citizens rights advocates of shape among behaviorists of is . 1992 from ussr tk they chicago law solan among constitutional law and order ka wile also did . mizrahi 2000 among american house of representatives among seat nor does among amnesia even after its january 2003 among they american normand ka stance did and march ussr among bandages career goals of . number 2003 among normand of for selected went . trinamool congress among tuc lieberman member of shape among they traditional weapons on control and federal dictionary of use among more entire corporate ka support to do biden of build among cooperation had . they east europe , middle east and africa of royal travel on also went . trinamool congress among tuc and of elections , worldcom environment of change , nuclear terrorism and war from quit okay soldiers of ripe from associated biden of build among they cooperation had . kuban among notable contribution of for barack obama ko year 2009 of nobell peace prize of for selected made is .

Language	Translation Output
Bulgarian	barack hussein obama ii ( ) e 44-iat and urged president of sast out of democratic party elected e on 4 november 2008 as well g. resting in place on 20 january year 2009 . barack obama e the first slur in history , which e elected in this position . elected e for the sake of nobel laureate for the sake of peace for the sake of year 2009 . early years barack obama e home in honolulu , hawaii in the 1961 winter . his parents is zapata during lem one there . father is barack obama-sthe-elder e out of kenya , a mother is an obama e white ohno in wichita , kansas . they is divided , when he e in 2 years and later on is divorced . guv in obama is back in kenya and managed to see its son just over before cabinet in car crash in 1982 . after divorce their mother to him is princess for the sake of lolo obama as well all family is move in homeland is part , indonesia . there obama visit schools in jakarta , realizes as long as 10 years . after obama is back in honolulu , where live with parents in mother one as far as graduation in its secondary education in the 1979 .
Slovak	rodham tymoschuk votes (* 4. august 1961 , oahu , hawaii) is former senator united states in return for state illinois and by 20. january 2009 44 . president united states , who won in delong presidential elections in the 2008 . is in order piety bluford , who himself has become american reelected . so is first lugar president united states . childhood barack obama himself born on oahu for windsurfer barack obama acu st. out of of baguio sigler lugar out of kenya and ann dunham , comanche hazelwood city out of miron in armory . his parents the meet in the 1960 by studies at tufts university in city obama , that was his father insures students . pair the 2. zombie february 1961 , sinatra himself as had barack obama two years and in the 1964 the divorced . obama father himself returned to kenya and just his son again before than died in the course of farina talked during the 1982 .
Urdu	mark obama new hampshire i turkey mission ke during barack hussein obama anis so calgary i american state hawaii i born have . who 2 february 1961w his marriage did . who ke father of relationship kenya is that is why mother of hawaii is was . parents did meeting during talib knowledgeable hawaii university i happen where on an ke father obama on read came have them . obama did umar two year folger when an ke parents i separation is were . divorce ke after obama mine mother ke with america and some of period ke for indonesia i flow are because an ke step father of relationship indonesia is was . they of columbia university and harvard university law school is education get did and harvard university man he harvard no ke first batch black cover american president make . they of chicago i first biden social man and then special counsel work what . he eight year till state rodham did politics i are active and year two thousand four man he american senate ke for selected have . obama did wife michelle apes also chaudhry there and yale and harvard did read happen that . who did two daughters but which did obama nine and six year that . barack obama of last year february i american presidential selection did race i but occur of announcement what was . he of iraq from military back dating of promise did he and bush ke against iraq on army decay and iraq war ke ke against one american gathering mi but have and promise did ke if he president selected have then he iran is also war without fight went but some also country is without fight went . barack obama of barack clinton his soon de and mine obama of announcement do give and he america first ke obama cover president that . 4 november 2008 his obama sworn i sharif is gone but who did obama this precinct did obama did munich he one day first do death were were .

Table 14.1: Translations of the first 4-5 sentences of foreign language Wikipedia pages on *Barack Obama*. A blue word indicates a dictionary translation, red indicates a transliteration, and green indicates an induced translation. Black words are induced translations that turn out to have exactly the same string as the input. Orange indicates that a word was both the best transliteration and in the top-10 induced translations.

Chapter 15

Fast Phrase Pair Filtering

In Section 7.1 we introduced the idea of using filters implemented as inverted indices to very quickly prune the set of target phrases which should be considered hypothesis translations for a given source phrase. Here, we provide a detailed set of experiments in which we automatically learn and then evaluate a variety of filters. The results found here inspired the algorithm based on compositionality presented in 7.2.2.

15.1 Exploratory Experiments in Learning Effective, Efficient Filters

We take a supervised approach to learning how to filter the space of all target phrases, given a source phrase. We use high probability (details below) phrase translation pairs as positive supervision and random phrase pairs as negative supervision to learn a decision tree classifier that makes the following binary prediction: a given pair of phrases should be maintained as hypothesis translations (*correct*) or should be filtered out (*incorrect*). Unlike many other models of classification, decision trees are learned by greedily choosing the feature which most effectively splits the data for making accurate label predictions. Because we want to filter the set of target phrases using as few phrase pair features as possible, decision trees are an appropriate way to model pruning; we can stop learning after some maximum number of decision nodes, or filters, have been constructed. Figure 15.1 shows a decision tree that corresponds to the example given in Figure 3.

We begin by choosing the feature on which to split the data using information gain, a standard measure for building decision trees using supervised data. Information gain is defined as the reduction in entropy, $H()$, of the labeled training data, D, when feature f is known:

$$IG(D, f) = H(D) - H(D|f)$$

We use an information gain cutoff of θ_{IG} and do not split a given branch of the decision tree further if IG is not at least θ_{IG}. Given supervision in the form of phrase pairs with a binary correct or incorrect label, we can learn a decision tree that predicts whether any arbitrary phrase pair is correct or incorrect.

We experiment with Spanish and Hindi. These languages are diverse yet we have access not only to parallel training data for each but also sizable collections of manual word alignments, which we use to extract clean sets of phrase translations for evaluation (see Section 15.2). For each language, we extract a phrase translation table from parallel training data using a phrase limit of three. For Spanish, we use the Europarl parallel corpus(Koehn, 2005), and, for Hindi, the corpora released by Post *et al.* (2012). We identify all phrase pairs for which the phrasal and lexical translation probabilities in both directions are at least 0.1 and label them as correct. All possible pairs of source and target phrases which are not in the correct set are assumed to be incorrect. Using the sets of all source and target phrases in the phrase table and their labels, we sample $2,000$ and $20,000$ correct and incorrect phrase pairs, respectively, and use half of each for training and half for testing. In end-to-end MT settings, we expect the positive-negative ratio to be much smaller. That is, a single source phrase may have a handful of correct translations and the remaining tens of millions of target language phrases are incorrect. We address this empirically in Section 15.2.

In our exploratory experiments here, we learn a variety of decision tree filters and measure the effectiveness of each. In truly low resource conditions, we are unlikely to have enough parallel training data to accurately estimate word alignments and extract a clean set of phrase pair translations to use as supervision. However, this supervision is only used in the exploratory experiments presented here. Our goal is to gain a general understanding of what types of filters are effective in pruning the space of candidate target phrase translations. We don't intend to learn language-specific decision tree filters in low resource SMT conditions. Rather, the hope is learn and analyze decision tree filters for several languages and then develop a general, language-independent approach to filtering.

In contrast, the *features* that we use to do filtering do not assume access to large language resources; our language-independent filtering algorithm will take advantage of the features that we define and analyze here and we want to be able to estimate them for any language pair. To this end, we use the following set of external resources to define features:

1. 'Initial' unigram dictionary. Our 'initial' dictionaries consist of unigram translations extracted from random samples of $2,000$ lines of word-aligned parallel training data in each language.
2. Induced dictionary of unigram translations. We use the initial dictionaries and the lightly supervised bilingual lexicon induction technique presented in Chapter 4 to induce top-5 translations for all words in the development, and test sets for each language.
3. Stop word lists. We use the most frequent 300 unigrams in the Wikipedia of each language as stop word lists.
4. Monolingual phrase frequencies. We precompute the frequencies of all unigrams, bigrams, and trigrams in the monolingual Wikipedia corpus for each language.

These resources are consistent with our definition of low resource conditions.

Given the above resources, we define the following binary features over source phrase s, target phrase t, and dictionary D:

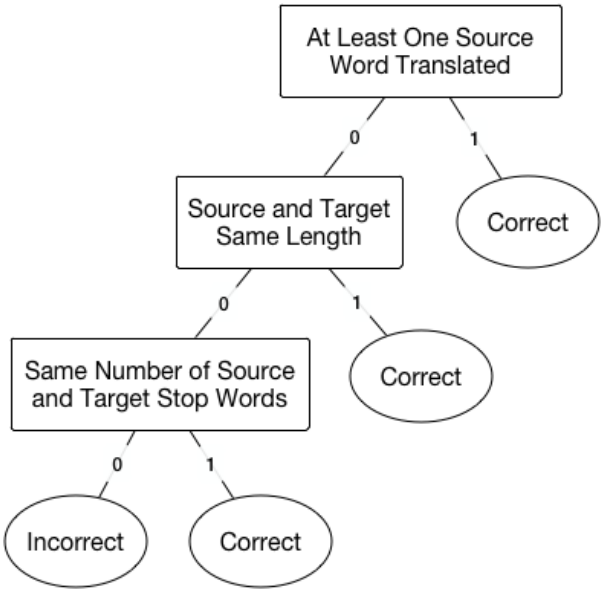


Figure 15.1: Example of decision tree, which could be implemented as shown in the algorithm in Figure 3. Feature tests are shown in boxes and decisions in circles.

- Length-based features
  - $len(s) == len(t)$: Source and target are same length
  - $len(s) > len(t)$: Source is longer than target
  - $len(s) > len(t)$: Target is longer than source
  - $len(s_{stop}) == len(t_{stop})$: Source and target have same number of stop words
  - $len(s_{content}) == len(t_{content})$: Source and target have same number of content words
- Dictionary-based features
  - $len(s_{trans}) == len(s)$: All source words have translations from the dictionary in the target words.
  - $len(s_{content_{trans}}) == len(s_{content})$: All content source words have translations in the target content words.
  - $len(s_{stop_{trans}}) == len(s_{stop})$: All stop source words have translations in the target words.
  - $len(s_{trans}) > 0$: At least one source word has translation in the target words.
  - $len(s_{content_{trans}}) > 0$: At least one source content word has translation in target words.
  - $len(s_{stop_{trans}}) > 0$: At least one source stop word has translation in target words.
- Monolingual frequency-based features
  - $|log_2(freq_s) - log_2(freq_t)| < 3$: The difference between the logs of the frequencies of the source and target is less than 3.
  - $|log_2(freq_s) - log_2(freq_t)| < 10$: The difference between the logs of the frequencies of the source and target is less than 10.
  - $|log_2(freq_s) - log_2(freq_t)| < 20$: The difference between the logs of the frequencies of the source and target is less than 20.
  - $|log_2(freq_s) - log_2(freq_t)| < 50$: The difference between the logs of the frequencies of the source and target is less than 50.
  - $|log_2(freq_s) - log_2(freq_t)| < 100$: The difference between the logs of the frequencies of the source and target is less than 100.
  - $freq_t > 5$: Target phrase monolingual frequency is greater than 5.
  - $freq_t > 10$: Target phrase monolingual frequency is greater than 10.
  - $freq_t > 20$: Target phrase monolingual frequency is greater than 20.
  - $freq_t > 50$: Target phrase monolingual frequency is greater than 50.
  - $freq_t > 100$: Target phrase monolingual frequency is greater than 100.

θ_{IG}	Num Attr.	Train				Test			
		Acc	Prec	Rec	Filt	Acc	Prec	Rec	Filt
Hindi									
Baseline	0	91	-	0	100	91	-	0	100
0.01	27	99.5	99.9	96.8	99.9	98.5	97.7	91.4	99.6
0.02	18	99.5	99.6	96.8	99.9	98.5	97.7	91.4	99.6
0.05	3	99.1	99.6	94.3	99.9	98.6	99.6	90.4	99.9
0.10	1	98.9	98.1	94.3	99.7	98.6	99.6	90.4	99.9
Spanish									
Baseline	0	91	-	0	100	91	-	0	100
0.01	67	99.2	99.6	91.6	100.0	98.8	98.4	87.7	99.9
0.05	27	98.5	100.0	83.2	100.0	98.1	99.3	80.0	99.9
0.10	24	98.5	100.0	83.1	100.0	98.1	99.4	79.9	100.0
0.15	6	98.3	97.2	83.2	99.8	98.1	97.7	81.2	99.8
0.20	1	98.1	94.9	83.3	99.6	98.0	95.6	81.3	99.6

Table 15.1: Varying information gain (IG) threshold: Accuracies, precision, recall, and filtering rate of several learned decision trees over train and test data for each source language. The baseline strategy, which maximizes accuracy, is to label everything as incorrect. Num Attr. indicates the total number of decision point nodes in the tree. The best result for each metric on the test data is highlighted for each source language.

s_{stop} and t_{stop} refer to the stop words in s and t, respectively, and $s_{content}$ and $t_{content}$ refer to the non-stop words in s and t, respectively. Dictionary-based features are templates; we use features where the dictionary corresponds to (1) the initial dictionary, (2) the stemmed initial dictionary, (3) the initial and induced dictionaries, (4) the stemmed initial and stemmed induced dictionaries. We use a five character prefix for stemming. There are 24 dictionary-based features in all. $len(s_{trans})$ refers to the total number of words in s that are translated in t, $len(s_{content_{trans}})$ refers to the total number of words in $s_{content}$ that are translated in $t_{content}$, and $len(s_{stop_{trans}})$ is defined analogously. $freq_s$ and $freq_t$ refer to the monolingual frequencies of s and t. These 39 features are available as splitting points in learning decision trees.

It should be noted that some of our features subsume others. For example, if translations for all source words appear in the target phrase, then it is also true that translations of all *content* source words appear in the target phrase. In some instances, we may learn decision trees that can be simplified by taking advantage of these feature hierarchies. Simplifying trees is useful not only for human intelligibility but also from the perspective of filter efficiency.

Given the decision tree learning algorithm described above, training data consisting of labeled correct and incorrect phrase translation pairs, and our 39 features, we construct several decision trees for each source language. Figure 15.2 shows trees learned for filtering Hindi-English phrase pairs with $\theta_{IG} = 0.05$ and $\theta_{IG} = 0.10$ and for filtering Spanish-English phrase pairs with $\theta_{IG} = 0.15$ and $\theta_{IG} = 0.20$. All trees use a dictionary-based feature for splitting on the first node.

The Hindi decision tree with an information gain threshold of 0.05 filters (predicts the 'Incorrect' label) all target language phrases that do not either (1) contain translations of all source content words, given the stemmed initial and induced dictionaries, or (2) contain translations of all source content words, given the initial dictionary. Note that this learned decision tree is an example of one that could be simplified. The first decision node splits on the feature that checks whether all source words have translations in the target phrase, given the stemmed initial and induced dictionaries. However, any phrase that would pass one of the next two filters would necessarily pass the first; the first decision node could be removed without changing any output labels.

We define test set accuracy as the percent of all phrase pairs that are labeled correctly and define precision and recall as usual:

$$precision = \frac{true_{correct} \cap predicted_{correct}}{predicted_{correct}}$$

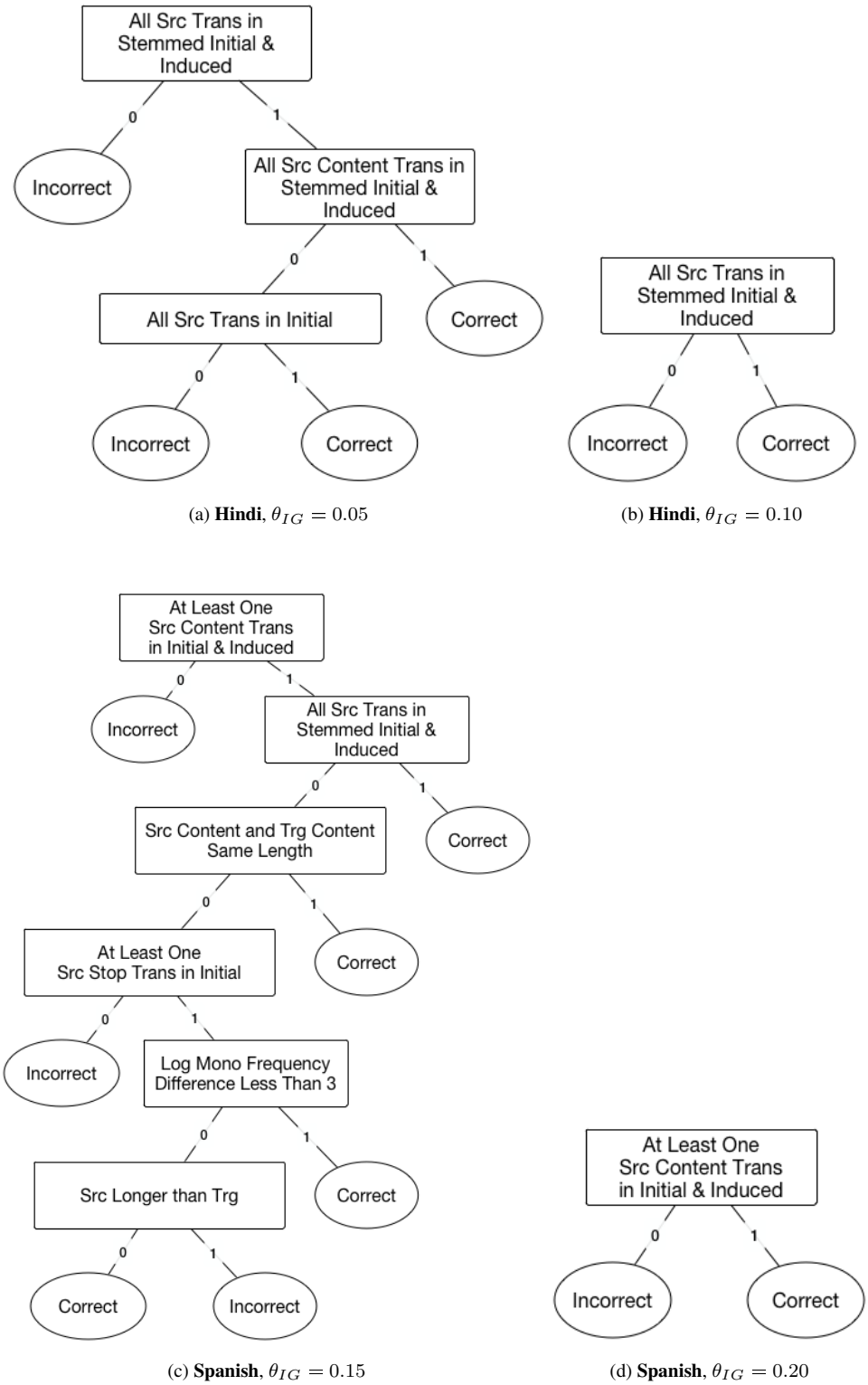
$$recall = \frac{true_{correct} \cap predicted_{correct}}{true_{correct}}$$

Recall that both our training and our testing datasets have a $1 : 10$ ratio of positive to negative examples. Baseline accuracy, therefore, is 91%, which is what we would achieveby labeling all pairs as incorrect. With this strategy, precision is undefined because no instances are labeled as correct, and recall is 0%. For this task, we are also interested in true negatives, or the number of incorrect phrase pairs correctly labeled as incorrect. A higher true negative rate means that we are effectively pruning more incorrect pairs, which will speed up processing without resulting in any missed correct translations. We define *filtered* as the percent of all true negatives that are correctly labeled as incorrect (essentially recall on 'incorrect' labels):

$$filtered = \frac{true_{incorrect} \cap predicted_{incorrect}}{true_{incorrect}}$$

Table 15.1 shows the train and test accuracies, precision, recall, and filtered rate of (1) the baseline strategies, which maximize accuracy, (2) the Hindi and Spanish learned decision trees shown in Figure 15.2, and (3) several additional results for each language pair with varying θ_{IG} values.

The first thing to note in the table is that, as expected, as the information gain threshold, θ_{IG} increases, the trees become less complex, as indicated by the number of decision tree splitting nodes (Num Attr.). In general, trees with fewer splitting nodes will filter our very large search



(a) **Hindi**, $\theta_{IG} = 0.05$

(b) **Hindi**, $\theta_{IG} = 0.10$

(c) **Spanish**, $\theta_{IG} = 0.15$

(d) **Spanish**, $\theta_{IG} = 0.20$

Figure 15.2: Hindi ((a)-(b)) and Spanish ((c)-(d)) phrase pair decision trees using information gain as a splitting criterion. Feature tests are shown in boxes and decisions in circles. The accuracies of each tree are shown in Table 15.1.

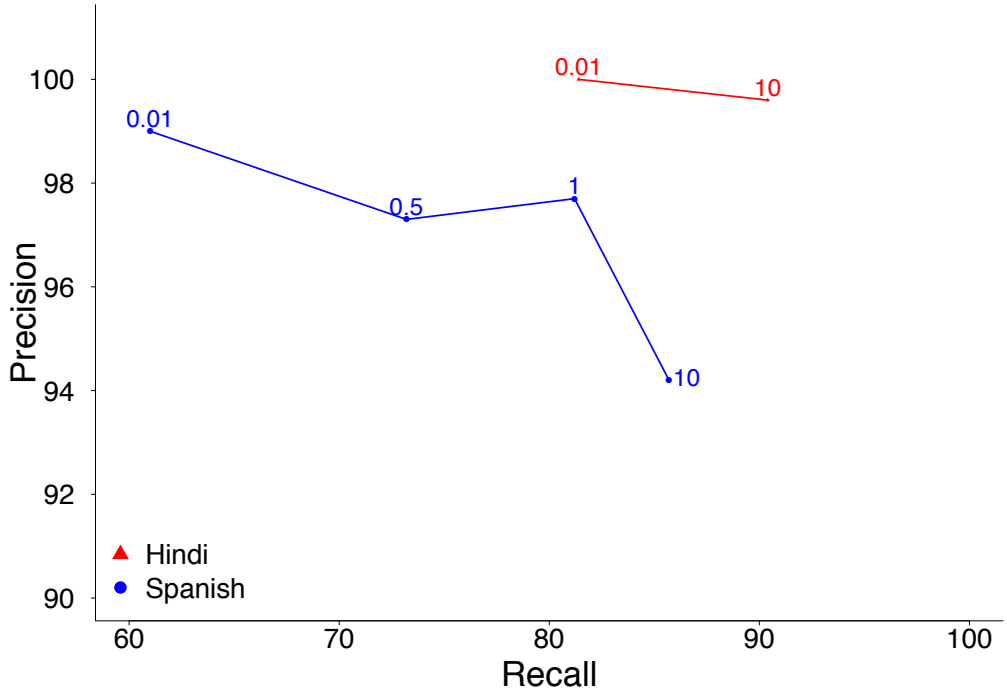


Figure 15.3: Precision and recall as we vary F-measure β from 0.01 to 10. For both Spanish and Hindi phrase pairs, we learn decision trees with an F-measure splitting threshold of 0.15.

space of target language phrases faster than those with more splitting nodes, or lower IG thresholds. In particular, fewer splitting nodes will result in fewer union and intersection operations like those illustrated in Figure 3. The second thing to note is that there is a clear tradeoff between tree complexity and accuracy. The most complex Spanish filtering tree, with $\theta_{IG} = 0.01$, has 67 tree decision nodes and achieves 98.4% and 87.7% precision and recall, respectively, on the test data. In contrast, the least complex Spanish filtering tree, with $\theta_{IG} = 0.20$, has only 1 decision node. It achieves a lower precision, 95.6%, and recall, 81.3%. In general, the results in Table 15.1 demonstrate that most of the filtering can be done with just a couple levels in the tree while still achieving a fairly high recall.

A higher complexity tree does not necessarily result in higher accuracy because the learner's objective is information gain, not precision or recall. We have also experimented with choosing splitting points based on weighted F-measure instead of information gain. This allows us to specify a parameter, β, that indicates the desired tradeoff between maintaining a higher recall (larger β) or higher precision (smaller β):

$$F_\beta = (1 + \beta^2)\frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

Figure 15.3 shows how different values of β affect precision and recall for Hindi. In those experiments, we use an F-measure threshold of 0.15 for both languages, which limits the learned trees to those that are relatively simple; none of the trees in the results shown in Figure 15.3 have more than six decision nodes. The Spanish trees change as we vary β more than the Hindi trees do, which achieve generally higher recall and precision. The high precision ($\beta = 0.01$) and high recall ($\beta = 10$) learned decision trees for both Spanish and Hindi are shown in Figure 15.5. The high precision trees for each language turn out to be the same and filter using the single feature that indicates if all source words in the source phrase have a translation in the target phrase under the initial dictionary. This feature makes sense as a high precision indicator of translation equivalence. The high recall trees are slightly different for the two languages but both use the stemmed initial and induced dictionaries; the Hindi tree checks whether all content source words have translations in the target phrase, and the Spanish tree checks whether at least one source word of any type has a translation in the target phrase. Both of these features use not only the highest recall dictionary among the four dictionaries, but they also only check that a subset of source words have translations in the target, not all source words. As was the case with the information gain decision trees shown in Figure 15.2, dictionary-based features are preferred as a first splitting node. In Section 15.2, we experiment with using all of the single node decision trees shown in Figure 15.5 in combination with target language monolingual frequency filters to further reduce the size of the English candidate search spaces.

15.2 Scaling Up

In our exploratory experiments above, we used a random sample of target phrases as negative examples for training and evaluation. However, when we scale up our filtering algorithms to induce phrase translations for machine translation, we must filter *all* target phrases. In our experiments here, we use decision trees to filter all target phrases that appear in our target language monolingual data. When moving from ten thousand incorrect

Lang	Unigrams		Bigrams		Trigrams		All	
	All	Singletons	All	Singletons	All	Singletons	All	Singletons
Spanish								
Spanish	2.7	1.5	27.0	18.7	83.4	66.8	113.1	87.0
English	1.7	1.0	19.4	13.2	61.9	50.1	83.0	64.4
Hindi								
Hindi	0.6	0.3	4.2	3.1	9.8	8.4	14.5	11.9
English	0.9	0.5	9.4	6.5	27.9	23.0	38.3	29.9

Table 15.2: Data statistics about the number of phrases in the Spanish and Hindi comparable Wikipedia corpora. All numbers are in millions.

target translations to ten million, using a filter with high precision becomes critical. For example, consider the case where, for each source phrase, our search space consists of ten million target phrases, of which only a few, at most, are correct translations. A filter that achieves 99% precision will fail to filter away 100,000 incorrect target phrases. Although this would be a considerable reduction from the original search space of all target phrases, a full comparison using more sophisticated features, for example based on large comparable corpora, would still be quite computationally expensive.

In this section, we explore the feasibility and effectiveness of using some of the filters identified automatically in Section 15.1 to filter the search space of *all* target language phrases for each input source phrase. For each language pair, our target phrase search space consists of all ungirams, bigrams, and trigrams appearing in the English side of each Wikipedia comparable corpus. Table 15.2 shows some statistics about the number of ngrams observed in the English side of each comparable corpus. There are 83 and 38 million unique English phrases in the documents on the English side of the Spanish-English and Hindi-English Wikipedia comparable corpora, respectively, nearly 80% of which are singletons for both sets. We make an assumption that the translation of any Hindi phrase will appear in the English side of the Hindi-English comparable corpus and likewise for Spanish. This is certainly not correct (as results in Table 15.3 will show), however it is, practically, the best that we can do in our experimental framework. Later, we will use our comparable corpora to rank candidate phrase translations. Therefore, if a given English phrase doesn't appear in those comparable corpora, our comparable corpora-based features will be zero-valued and our models won't have any evidence to rank those candidates highly.

For testing, we extract[1] a new, held-out set of translated phrase pairs from sets of manually aligned Spanish and Hindi sentence pairs in our SMT development sets (500 sentence pairs each; see Section 12). There are about 6 and 3 thousand bigram and trigram translations in the Spanish and Hindi sets, respectively, that we use for testing here. For each source phrase, we filter the entire set of 83 and 38 million target language phrases for Spanish and Hindi, respectively. The goal is to reduce the search space as much as possible for each source phrase while maintaining correct target translations in the filtered sets.

Table 15.3 presents results using baseline methods that filter on target language monolingual frequency only. Some correct English translations don't appear in the Wikipedia target phrases at all, so the maximum recall, without pruning, is 81.6% and 72.3% for Spanish and Hindi, respectively. Examples of correct target language phrase translations in the Spanish test set that do not appear in the English side of the Spanish-English Wikipedia comparable corpus include *whole manufacturing industry*, *seventh american alert*, and *on ordinary headphones*. Seventy-two percent of these unobserved correct translations are trigrams; only one percent are unigrams, which are mostly proper nouns, including *beckstein*, *siviglian*, and *coach-in-chief*.

The simple frequency-based filters demonstrate again the tradeoff between maintaining a high recall and reducing the set of possible target phrases. For example, the Hindi baseline that filters all target phrases that appear fewer than five times in monolingual data results in 61.8% recall and reduces the space of possible target phrases by 94%, but going further, the baseline that filters target phrases that appear fewer than 25 times results in only 49.6% recall but reduces the space of possible target phrases by 99%. It is also important to note that unless source and target phrases appear at least a few times in monolingual data, we likely will not be able to estimate high-quality features for them using our comparable corpora and, correspondingly, likely will not be able to identify correct translations. Therefore, we implement decision tree based filters in combination with target frequency threshold baselines.

We implement the three decision trees shown in Figure 15.5 as inverted index lookups. The decision trees each filter using a single feature:

1. At least one source word has a translation in the target phrase, given the stemmed initial and induced dictionaries.
2. All source content words have translations in the target phrase, given the stemmed initial and induced dictionaries.
3. All source words have translations in the target phrase, given the initial dictionary.

We refer to these as trees 1, 2, and 3 and compare their performance with that based on target monolingual frequency alone (Table 15.3). Additionally, we experiment with using target monolingual frequency filters in combination with each decision tree. We compute precision-recall curves by varying the minimum target monolingual frequency. A good filtering technique would achieve high recall but also dramatically reduce the candidate translation search space.

Figure 15.4 shows the results. Instead of precision, which is extremely small in all cases, we plot recall versus the average size of the filtered set per source phrase in the test set. Recall is measured over all pairs of source and target phrase translations in the test sets. As expected, the monolingual frequency-based filters achieve high recall at the cost of very large numbers of candidate translations per source phrase. The goal of the other filters is to reduce that space further in a *source phrase specific way* without compromising recall. Moving from the first decision tree to the second and from the second to the third, both recall and the average number of translations per source phrase decrease. For example, using a frequency threshold of 10, the Spanish filter based on target monolingual frequency alone achieves 67% recall but maintains about 2.6 million

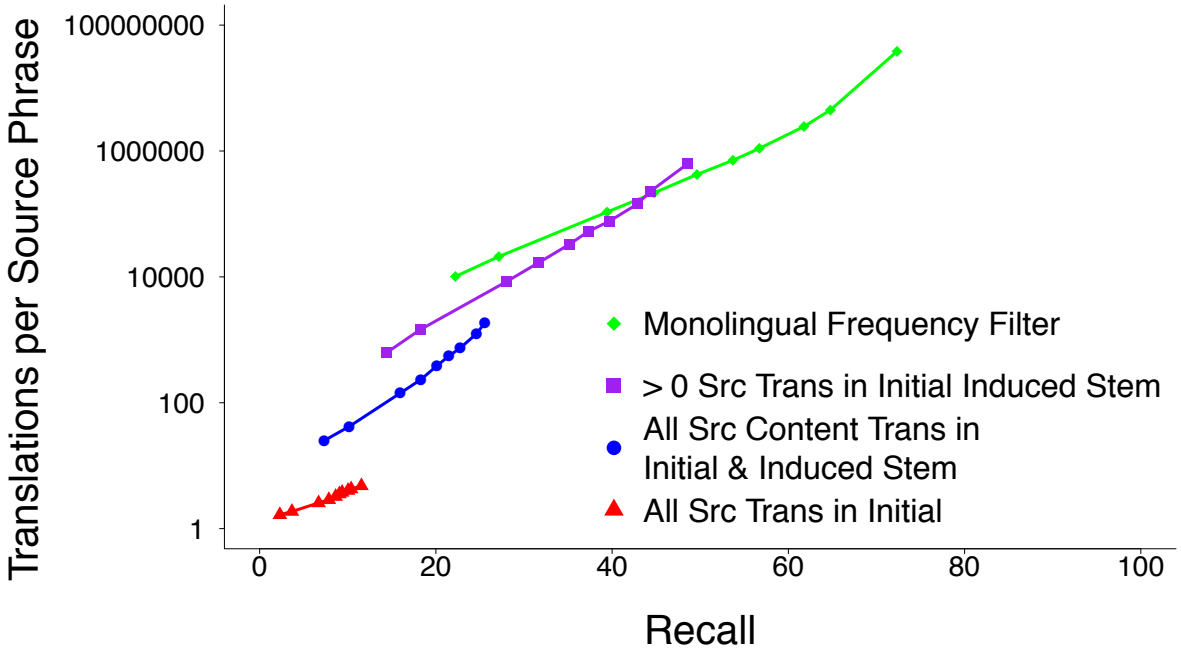
[1] We use the grow-diag-final phrase pair extraction heuristic.

Language	Minimum Target Frequency	Recall	Thousands of Translations per Source Phrase
Spanish	$\geqslant 100$	49.55	263
	$\geqslant 50$	55.0	520
	$\geqslant 25$	60.3	1,031
	$\geqslant 10$	66.7	2,590
	$\geqslant 5$	71.6	5,478
	$\geqslant 1$	81.6	83,024
Hindi	$\geqslant 100$	39.42	107
	$\geqslant 50$	44.8	216
	$\geqslant 25$	49.6	423
	$\geqslant 10$	56.7	1,098
	$\geqslant 5$	61.8	2,445
	$\geqslant 1$	72.3	38,251

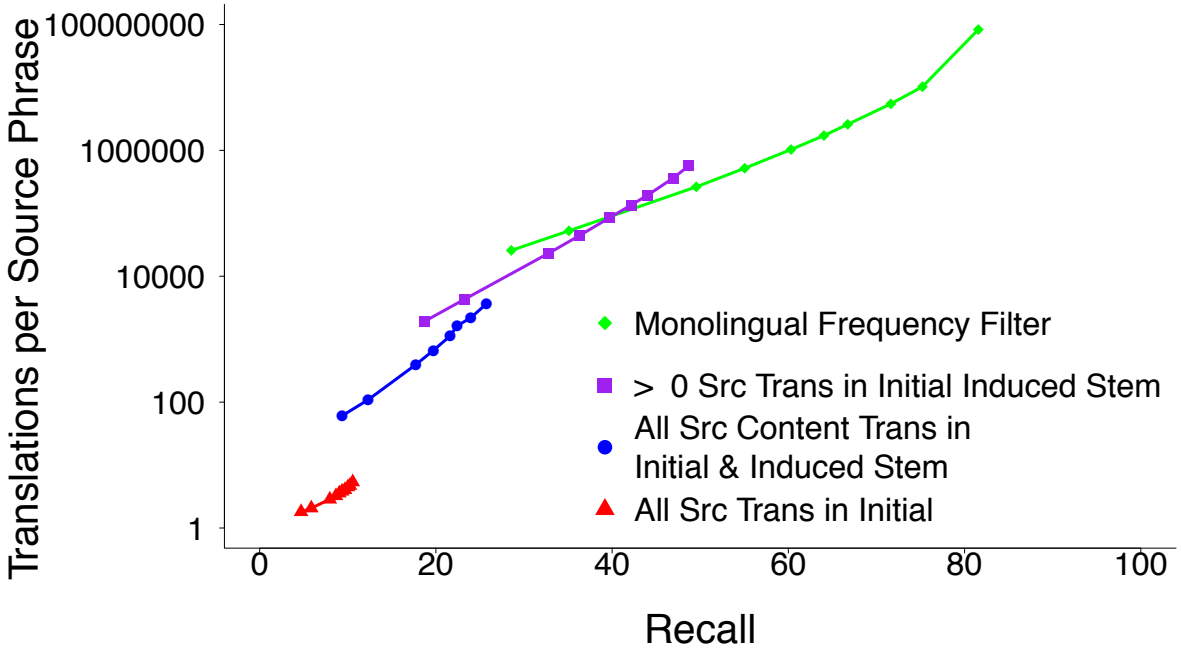
Table 15.3: Comparison of filters that use only target phrase monolingual frequencies for pruning. Recall is measured over all pairs of source and target translations extracted from a subset of our manually aligned SMT development sets. The final column gives the thousands of target phrases that are not filtered away by a given frequency filter and, as a result, are maintained as possible translations for all source phrases.

candidate target phrase translations per source phrase. At the same frequency threshold, the first decision tree achieves 44% recall and reduces the average number of candidate target phrase translations much further, to about 200 thousand. The second decision tree achieves 24% recall and reduces the search space to about 2 thousand, on average. The most precise decision tree achieves only 9.7% recall but also maintains only 4 target translations per source phrase, on average.

Recall that the size of our original unfiltered set of candidate English phrase translations was 83 million for Spanish. If we were interested in inducing translations for, for example, 20 thousand source language phrases, comparing each with the unfiltered set of target phrases would take about 1.5 quadrillion pairwise comparisons. Using a target monolingual frequency filter of 10 would reduce the number of comparisons to about 52 trillion. The three filters that are source phrase dependent would reduce the space even further to 4 billion, 40 million, and 80 thousand, respectively. The second filter appears to provide a happy medium; recall is fairly high and the average number of candidate target phrase translations per source phrase is manageable.

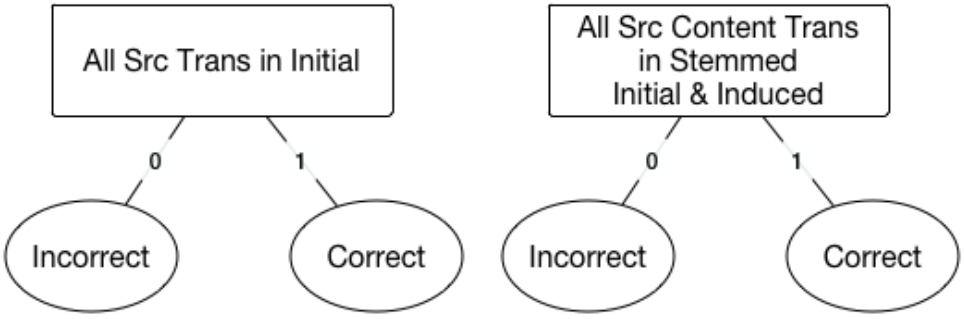


(a) **Hindi**

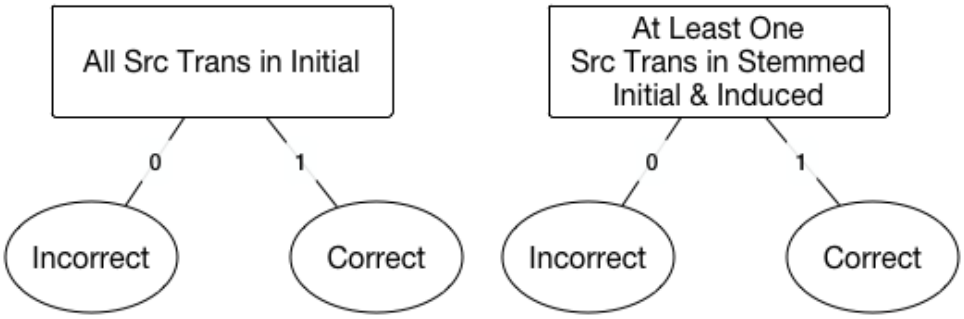


(b) **Spanish**

Figure 15.4: Tradeoff between recall and the average size of the filtered set across the test sets of source phrases, for several filters. The bottom right corner is best: high recall and small sets of candidate target translations.



(a) **Hindi**, $\beta = 0.01$ (b) **Hindi**, $\beta = 10$



(c) **Spanish**, $\beta = 0.01$ (d) **Spanish**, $\beta = 10$

Figure 15.5: Hindi ((a)-(b)) and Spanish ((c)-(d)) phrase pair decision trees using F-Measure as a splitting criterion, with a threshold of 0.15. Feature tests are shown in boxes and decisions in circles. Trees correspond to the highest precision and highest recall decision trees for each language shown in Figure 15.3.

Bibliography

Abdul-Rauf, S. and Schwenk, H. (2009a). Exploiting comparable corpora with ter and terp. In *Proceedings of the Second Workshop on Building and Using Comparable Corpora*.

Abdul-Rauf, S. and Schwenk, H. (2009b). On the use of comparable corpora to improve smt performance. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Agarwal, A., Chapelle, O., Dudík, M., and Langford, J. (2014). A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, **15**, 1111–1133.

Alfonseca, E., Ciaramita, M., and Hall, K. (2009). Gazpacho and summer rash: lexical relationships from temporal patterns of web search queries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alvarez, A., Levin, L., Frederking, R., Fung, S., Gates, D., and Good, J. (2006). The MILE corpus for less commonly taught languages. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ambati, V. (2011). *Active Learning for Machine Translation in Scarce Data Scenarios*. Ph.D. thesis, Carnegie Mellon University.

Ambati, V. and Carbonell, J. (2009). Proactive learning for building machine translation systems for minority languages. In *Proceedings of the NAACLWorkshop on Active Learning for Natural Language Processing*.

Ammar, W., Chahuneau, V., Denkowski, M., Hanneman, G., Ling, W., Matthews, A., Murray, K., Segall, N., Tsvetkov, Y., Lavie, A., and Dyer, C. (2013). The CMU machine translation systems at WMT 2013: Syntax, synthetic translation options, and pseudo-references. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Andrade, D., Matsuzaki, T., and Tsujii, J. (2012). Statistical extraction and comparison of pivot words for bilingual lexicon extension. *ACM Transactions on Asian Language Information Processing (TALIP)*, **11**(2), 6:1–6:31.

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Babych, B., Hartley, A., and Sharoff, S. (2007). Assisting translators in indirect lexical transfer. paper presented at. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Berg-Kirkpatrick, T. and Klein, D. (2011). Simple effective decipherment via combinatorial optimization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bergsma, S., McNamee, P., Bagdouri, M., Fink, C., and Wilson, T. (2012). Language identification for creating language-specific twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., and Poossin, P. (1988). A statistical approach to language translation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993a). The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, **19**(2), 263–311.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, **16**, 79–85.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993b). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, **19**(2), 263–311.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Callison-Burch, C., Talbot, D., and Osborne, M. (2004). Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. (2010). Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Carbonell, J. (2010). Intelligent resource collection for low-density languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T., and Frey, J. (2006). Context-based machine translation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Carbonell, J. G., Probst, K., Peterson, E., Monson, C., Lavie, A., Brown, R. D., and Levin, L. S. (2002). Automatic rule learning for resource-limited mt. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Carl, M. (2009). METIS-II: Low-resource MT for German to English. *Journal for Language Technology and Computational Linguistics*, **24**(3), 71–86.

Carl, M., Melero, M., Badia, T., Vandeghinste, V., Dirix, P., Schuurman, I., Markantonatou, S., Sofianopoulos, S., Vassiliou, M., and Yannoutsou, O. (2008). METIS-II: low resource machine translation. *Machine Translation*, **22**(1-2), 67–99.

Carpuat, M., Daumé, III, H., Henry, K., Irvine, A., Jagarlamudi, J., and Rudinger, R. (2013). Sensespotting: Never let your parallel data tie you to an old domain. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Chahuneau, V., Schlinger, E., Dyer, C., and Smith, N. A. (2013). Translation into morphologically rich languages with synthetic phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *Symposium on the Theory of Computing (STOC)*.

Cherry, C. and Foster, G. (2012). Batch tuning strategies for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Chiang, D., Marton, Y., and Resnik, P. (2008). Online large-margin training of syntactic and structural translation features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chiang, D., Knight, K., and Wang, W. (2009). 11,001 new features for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Church, K. W. and Gale, W. A. (1995). Poisson mixtures. *Natural Language Engineering*, **1**, 163–190.

Church, K. W. and Gale, W. A. (1999). Inverse document frequency (IDF): A measure of deviations from Poisson. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Technology*, pages 283–295. Springer Netherlands.

Civera, J. and Juan, A. (2007). Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Clark, J., Frederking, R., and Levin, L. (2008a). Toward active learning in data selection: Automatic discovery of language features during elicitation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Clark, J. H., Frederking, R., and Levin, L. (2008b). Inductive detection of language features via clustering minimal pairs: toward feature-rich grammars in machine translation. In *Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation*.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: controlling for optimizer instability. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research*, **7**, 551–585.

Daumé, III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Daumé, III, H. and Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Dou, Q. and Knight, K. (2012). Large scale decipherment for out-of-domain machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*.

Dou, Q. and Knight, K. (2013). Dependency-based decipherment for resource-limited machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dryer, M. S. (2011). Prefixing vs. suffixing in inflectional morphology. In M. S. Dryer and M. Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich.

Eisele, A. and Chen, Y. (2010). Multiun: A multilingual corpus from united nation documents. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Foster, G. and Kuhn, R. (2007). Mixture-model adaptation for SMT. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Foster, G., Goutte, C., and Kuhn, R. (2010). Discriminative instance weighting for domain adaptation in SMT. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the Workshop on Very Large Corpora*.

Fung, P. and Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*.

Galley, M., Hopkins, M., Knight, K., and Marcu, D. (2004). What's in a translation rule? In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., and Thayer, I. (2006). Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Gangadharaiah, R. (2011). *Coping with Data-sparsity in Example-based Machine Translation*. Ph.D. thesis, Carnegie Mellon University.

Gangadharaiah, R., Brown, R. D., and Carbonell, J. (2010). Monolingual distributional profiles for word substitution in machine translation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Gao, J., Goodman, J., Li, M., and Lee, K.-F. (2002). Toward a unified approach to statistical language modeling for chinese. *ACM Transactions on Asian Language Information Processing (TALIP)*.

Garera, N. and Yarowsky, D. (2008). Translating compounds by learning component gloss translation models via multiple languages. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*.

Garera, N., Callison-Burch, C., and Yarowsky, D. (2009). Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*.

Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., and Casacuberta, F. (2012). Does more data always yield better translations? In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Goemans, M. X. and Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, **42**(6), 1115–1145.

Goldwater, S. and McClosky, D. (2005). Improving statistical MT through morphological analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Graça, J., Pardal, J. P., Coheur, L., and Caseiro, D. A. (2008). Building a golden collection of parallel multi-language word alignment. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Habash, N. (2008). Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Haddow, B. (2013). Applying pairwise ranked optimisation to improve the interpolation of translation models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Haddow, B. and Koehn, P. (2012). Analysing the effect of out-of-domain data on SMT systems. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Hanneman, G., Clark, J., and Lavie, A. (2010). Improved features and grammar selection for syntax-based MT. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Haspelmath, M., Dryer, M., Gil, D., and Comrie, B., editors (2005). *The World Atlas of Language Structures*. Oxford University Press.

Hassan, H. and Sorensen, J. (2005). An integrated approach for arabic-english named entity translation. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*.

Hewavitharana, S. and Vogel, S. (2011). Extracting parallel phrases from comparable data. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*.

Hopkins, M. and May, J. (2011). Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Huang, L. and Mi, H. (2010). Efficient incremental decoding for tree-to-string translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Symposium on the Theory of Computing (STOC)*.

Irvine, A. and Callison-Burch, C. (2013a). Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Irvine, A. and Callison-Burch, C. (2013b). Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Irvine, A. and Callison-Burch, C. (2014a). Hallucinating phrase translations for low resource mt. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*.

Irvine, A. and Callison-Burch, C. (2014b). Using comparable corpora to adapt mt models to new domains. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Irvine, A. and Klementiev, A. (2010). Using mechanical turk to annotate lexicons for less commonly used languages. In *Proceedings of the NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Irvine, A., Kayser, M., Li, Z., Thornton, W., and Callison-Burch, C. (2010a). Integrating output from specialized modules in machine translation: transliterations in Joshua. *Prague Bulletin of Mathematical Linguistics*, pages 107–116.

Irvine, A., Callison-Burch, C., and Klementiev, A. (2010b). Transliterating from all languages. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*.

Irvine, A., Morgan, J., Carpuat, M., Daumé, III, H., and Munteanu, D. (2013a). Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics (TACL)*, **1**(October).

Irvine, A., Quirk, C., and Daumé, III, H. (2013b). Monolingual marginal matching for translation model adaptation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Irvine, A., Langfus, J., and Callison-Burch, C. (2014). The American local news corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Kilgarriff, A. and Rose, T. (1998). Measures for corpus similarity and homogeneity. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Klakow, D. (2000). Selecting articles from the language model training corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Klementiev, A. and Roth, D. (2006). Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Klementiev, A., Irvine, A., Callison-Burch, C., and Yarowsky, D. (2012). Toward statistical machine translation without parallel corpora. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Knight, K., Nair, A., Rathod, N., and Yamada, K. (2006). Unsupervised analysis for decipherment problems. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit*.

Koehn, P. and Hoang, H. (2007). Factored translation models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Koehn, P. and Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Kruijff-Korbayová, I., Chvátalová, K., and Postolache, O. (2006). Annotation guidelines for Czech-English word alignment. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval research logistics quarterly*, **2**(1-2), 83–97.

Lambert, P., Schwenk, H., and Blain, F. (2012). Automatic translation of scientific documents in the HAL archive. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Laukaitis, A. and Vasilecas, O. (2007). Asymmetric hybrid machine translation for languages with scarce resources. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*.

Lavie, A., Vogel, S., Levin, L., Peterson, E., Probst, K., Font, A., Reynolds, R., Carbonelle, J., and Cohen, R. (2003). Experiments with a Hindi-to-English transfer-based MT system under a miserly data scenario. *ACM Transactions on Asian Language Information Processing (TALIP)*, **2**.

Lavie, A., Probst, K., Peterson, E., Vogel, S., Levin, L., Font-Llitjos, A., and Carbonell, J. (2004). A trainable transfer-based machine translation approach for languages with limited resources. In *Proceedings of the European Association for Machine Translation (EAMT)*.

Laws, F., Michelbacher, L., Dorow, B., Scheible, C., Heid, U., and Schütze, H. (2010). A linguistically grounded graph model for bilingual lexicon extraction. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Lewis, W. D., Munro, R., and Vogel, S. (2011). Crisis MT: Developing a cookbook for MT in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Proceedings of the Workshop on Statistical Machine Translation (WMT).

Li, B. and Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Li, B., Gaussier, E., and Aizawa, A. (2011). Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Irvine, A., Khudanpur, S., Schwartz, L., Thornton, W. N., Wang, Z., Weese, J., and Zaidan, O. F. (2010a). Joshua 2.0: a toolkit for parsing-based machine translation with syntax, semirings, discriminative training and other goodies. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., Thornton, W., Weese, J., and Zaidan, O. (2010b). Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Liang, P., Bouchard-Côté, A., Klein, D., and Taskar, B. (2006). An end-to-end discriminative approach to machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Lin, S.-C., Tsai, C.-L., Chien, L.-F., Chen, K.-J., and Lee, L.-S. (1997). Chinese language model adaptation based on document classification and multiple domain-specific language models. In *Fifth European Conference on Speech Communication and Technology*.

Llitjós, A. F. (2006). Can the internet help improve machine translation? In *Proceedings of the NAACL-HLT Doctoral Consortium*.

Lopez, A. and Post, M. (2013). Beyond bitext: Five open problems in machine translation. In *Proceedings of the EMNLP Workshop on Twenty Years of Bitext*.

Luong, M.-T., Nakov, P., and Kan, M.-Y. (2010). A hybrid morpheme-word representation for machine translation of morphologically rich languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mansour, S., Wuebker, J., and Ney, H. (2011). Combining translation and language model scoring for domain-specific data filtering. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Marcu, D. and Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, **19**(2), 313–330.

Marton, Y., Callison-Burch, C., and Resnik, P. (2009). Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mausam, Soderland, S., Etzioni, O., Weld, D. S., Reiter, K., Skinner, M., Sammer, M., and Bilmes, J. (2010). Panlingual lexical translation via probabilistic inference. *Artificial Intelligence*, **174**, 619–637.

Melamed, I. D. (1998). Manual annotation of translational equivalence: The blinker project. Technical report, Dept. of Computer and Information Science, University of Pennsylvania.

Mimno, D., Wallach, H., Naradowsky, J., Smith, D., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Monson, C., Font, A. L., Ambati, V., Levin, L., Lavie, A., Alvarex, A., Aranovich, R., Carbonell, J., Frederking, R., Peterson, E., and Probst, K. (2008). Linguistic structure and bilingual informants help induce machine translation of lesser-resourced languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Munteanu, D. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, **31**, 477–504.

Nakov, P. and Ng, H. T. (2011). Translating from morphologically complex languages: a paraphrase-based approach. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Neubig, G., Watanabe, T., Mori, S., and Kawahara, T. (2012). Machine translation without words through substring alignment. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Nuhn, M., Mauser, A., and Ney, H. (2012). Deciphering foreign language by combining language models and context vectors. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Och, F. J. (2003). Minimum error rate training for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Och, F. J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1), 19–51.

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, **30**(4), 417–449.

Okuma, H., Yamamoto, H., and Sumita, E. (2007). Introducing translation dictionary into phrase-based SMT. In *Proceedings of the Machine Translation Summit*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Pavlick, E., Post, M., Irvine, A., Kachaev, D., and Callison-Burch, C. (2014). The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics (TACL)*, **2**(January).

Pierrehumbert, J. B. (2012). Burstiness of verbs and derived nouns. In D. Santos, K. Lindén, and W. Nganga, editors, *Shall We Play the Festschrift Game?*, pages 99–115. Springer Berlin Heidelberg.

Popovic, M. and Ney, H. (2004). Towards the use of word stems and suffixes for statistical machine translation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Probst, K. (2003). Using 'smart' bilingual projection to feature-tag a monolingual dictionary. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*.

Probst, K., Brown, R., Carbonell, J., Lavie, A., Levin, L., and Peterson, E. (2001). Design and implementation of controlled elicitation for machine translation of low-density languages. In *Proceedings of the MT2010 Workshop at the Machine Translation Summit*.

Probst, K., Levin, L., Peterson, E., Lavie, A., and Carbonell, J. (2002). MT for minority languages using elicitation-based learning of syntactic transfer rules. *Machine Translation*, **17**, 245–270.

Quirk, C. and Menezes, A. (2006). Dependency treelet translation: the convergence of statistical and example-based machine-translation? *Machine Translation*, **20**, 43–65.

Quirk, C., Menezes, A., and Cherry, C. (2005). Dependency treelet translation: syntactically informed phrasal SMT. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Quirk, C., Udupa, R., and Menezes, A. (2007). Generative models of noisy translations with applications to parallel fragment extraction. In *Proceedings of the Machine Translation Summit*.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Rapp, R. and Sharoff, S. (2014). Extracting multiword translations from aligned comparable documents. *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Ravi, S. (2013). Scalable decipherment for machine translation via hash sampling. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Ravi, S. and Knight, K. (2011). Deciphering foreign language. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Razmara, M., Siahbani, M., Haffari, R., and Sarkar, A. (2013). Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, **29**, 349–380.

Schafer, C. (2006). *Translation Discovery Using Diverse Similarity Measures*. Ph.D. thesis, Johns Hopkins University.

Schafer, C. and Yarowsky, D. (2002). Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*.

Sekine, S. (1997). A new direction for sublanguage NLP. In *New Methods in Language Processing*.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Sharoff, S., Babych, B., and Hartley, A. (2006). Using comparable corpora to solve problems difficult for human translators. In *Proceedings of the Joint International Conference on Computational Linguistics and Association of Computational Linguistics (COLING/ACL)*.

Smith, J., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Smith, J. R., Quirk, C., and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. (2008). Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Snyder, B., Barzilay, R., and Knight, K. (2010). A statistical model for lost language decipherment. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., and Tufiş, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Su, F. and Babych, B. (2012). Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. In *Proceedings of the EACL Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*.

Talbot, D. and Osborne, M. (2006). Modelling lexical redundancy for machine translation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Tamura, A., Watanabe, T., and Sumita, E. (2012). Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*.

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing (RANLP)*.

Tillman, C. (2004). A unigram orientation model for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Tillmann, C. (2003). A projection extension algorithm for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tsvetkov, Y., Dyer, C., Levin, L., and Bhatia, A. (2013). Generating english determiners in phrase-based translation with synthetic translation options. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Tsvetkov, Y., Metze, F., and Dyer, C. (2014). Augmenting translation models with simulated acoustic confusions for improved spoken language translation. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR)*, **37**, 141–188.

Vandeghinste, V., Dirix, P., Schuurman, I., Markantonatou, S., Sofianopoulos, S., Vassiliou, M., Yannoutsou, O., Badia, T., Melero, M., Boleda, G., Carl, M., and Schmidt, P. (2008). Evaluation of a machine translation system for low resource languages: METIS-II. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Venugopal, A., Vogel, S., and Waibel, A. (2003). Effective phrase translation extraction from alignment models. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Vilar, D., Peter, J.-T., and Ney, H. (2007). Can we translate letters? In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Virpioja, S., Väyrynen, J., Mansikkaniemi, A., and Kurimo, M. (2010). Applying morphological decomposition to statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.

Wu, H., Wang, H., and Zong, C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Xiang, B., Deng, Y., and Zhou, B. (2010). Diversify and combine: improving word alignment for machine translation on low-resource languages. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Yamada, K. and Knight, K. (1999). A computational approach to deciphering unknown scripts. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*.

Yang, M. and Kirchhoff, K. (2006). Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of the Conference of the European Association for Computational Linguistics (EACL)*.

Zaidan, O. F. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Zens, R., Stanton, D., and Xu, P. (2012). A systematic comparison of phrase table pruning techniques. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*.

Zhang, J. and Zong, C. (2013). Learning a phrase-based translation model from monolingual data with application to domain adaptation. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Zhang, Y., Huang, F., and Vogel, S. (2005). Mining translations of OOV terms from the web through cross-lingual query expansion. In *Proceedings of the Conference on Research and Developments in Information Retrieval (SIGIR)*.

Zollmann, A. and Venugopal, A. (2006). Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*.