LEARNING FORMALITY FROM JAPANESE-ENGLISH PARALLEL CORPORA

Liam Dugan

A THESIS

in

Robotics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Master of Robotics

2020

Supervisor of Thesis

Chris Callison-Burch, Associate Professor of Computer and Information Science

Supervisor of Thesis

Camillo Jose Taylor, Professor of Computer and Information Science

Graduate Group Chairperson

M. Ani Hsieh, Research Associate Professor of Mechanical Engineering

LEARNING FORMALITY FROM JAPANESE-ENGLISH PARALLEL CORPORA

Dedicated to my family. They have been my rock during quarantine.

Without them, none of this would have been possible.

ACKNOWLEDGEMENT

I would like to start by thanking the members of the original CIS530 Semi-Formal team: Trevor Huang, Gianluca Gross, and Will Lowe, for believing in me enough to agree to tackle the original version of this project with me. Your faith was much appreciated at the time and still is today.

I would also like to thank the members of Dr. Chris Callison-Burch's lab for their countless hours spent answering questions and listening to me rant about this project. In particular, Li Zhang was instrumental in giving feedback on the first two chapters, Reno Kriz helped very early on with getting models running and later provided crucial editing support at the latest of hours, and Daphne Ippolito served as an invaluable source of advice throughout this project.

I'd also like to thank my roommates Ziad Ben Hadj-Alouane and Henry Zhu for keeping me company during my final year and always encouraging me to pursue my research goals no matter how tough the road may seem.

I want to extend a special thanks to the rest of the members of the Science and Technology Wing Class of 2019 (et al.): Sadat Shaik, Jared Winograd, Zhengyi Luo, Bradley Jackson, Emmett Neyman, Xuerui Fa, Jason Schwartz, Yueqi Ren, Micah Getz, Tyler Durkin, Kyle Kersey, and Sander Depietri. You all are the reason I maintained my sanity in undergrad and you continue to fulfill this role during quarantine. Your friendship is truly a gift and I treasure it every day.

Last, but not least, I'd like to express my deepest gratitude to Dr. Chris Callison-Burch for continually going out of his way to ensure that this project is done well. From his initial vote of confidence, to his generous financial support, to his insightful feedback on my writing, he has continued to treat my research with a level of respect that I am doubtful it will ever deserve. This project would not be what it is today without his support.

ABSTRACT

LEARNING FORMALITY FROM JAPANESE-ENGLISH PARALLEL CORPORA

Liam Dugan

Chris Callison-Burch

Machines that can automatically classify a sentence as either "Formal" or "Informal" are known as Formality Classifiers. These classifiers are broadly useful in many applications. For example, they can notify a user that an email they've written is of unusually low formality, or they can be used to automatically rank the dialogue options of a system such as Amazon's Alexa or Apple's Siri to ensure proper formal communication.

Traditionally, English formality classifiers have been trained using supervised data, meaning that they use English sentences with manually annotated formality labels. In this work, we demonstrate how we can accurately predict formality without requiring full supervision. To do this, we leverage a Japanese-English parallel corpus, relying on the fact that Japanese verbs contain formality markers, unlike English. We create a formality dataset consisting of over one million automatically labeled English sentences, and show that a classifier trained on our data outperforms those trained on previous manually labeled formality datasets.

In doing so, we raise questions about the suitability of current supervised data sources for proper formality evaluation and claim that current sources of this data contain a significant topic bias. We claim that formality is encapsulated entirely within the relationship between two speakers and argue against the inclusion of topic as a feature for formality classifiers.

Finally, we offer a proof of concept study for the applicability of adversarial decomposition techniques to train topic-agnostic formality classifiers. We show that, in corpora with minimal topical bias, adversarially decomposed representations of formality achieve promising results in classification.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

While the phrases "What's up?" and "How are you?" are similar in meaning, the differ greatly in terms of their formality. The ability to automatically detect which of these two is the formal expression would be broadly useful in many applications. The automatic detection of formality is called the formality classification task.

There are three varieties of the formality classification task. The first is the lexical (or word-level) formality classification task. That is, given a single word, classify it as either formal or informal.

The second variety is the sentence-level formality classification task. That is, given a sentence (in isolation), classify it as either formal or informal.

The final variety is the document-level formality classification task. This task is similar to the previous two. Given a document, classify it as either formal or informal.

In this work we will be attempting to improve sentence-level formality classification.

Traditionally, machines that perform the sentence-level formality classification task (a.k.a. classifiers) have been trained using supervised data. This means that they are given hundreds or thousands of English sentences with manually annotated formality labels. After being given this data, classifiers use the labeled sentences to extract patterns that are common to all sentences in a specific class (i.e. "they train on the dataset"). Classifiers can then use a combination of these patterns to dynamically infer the class of new unseen sentences.

In order to improve these classifiers, there are two main routes. 1) improve the algorithm's ability to detect patterns, or 2) gather more training data. We will be attempting to do the latter.

In order to gather more data for training, there are typically two options. 1) pay human annotators to manually label new sentences for formality (creating more supervised data),

or 2) find a way to automatically infer the label of a sentence using some extra outside information (creating semi-supervised data or unsupervised data). In this work we will be attempting to gather more data in this second way.

The source of outside formality information we will use to build our dataset will be a Japanese-English parallel corpus. This is a large collection of English and Japanese sentences that are translations of one another. We will use the fact that formality is explicitly encoded in Japanese verb endings as our source of labels and we will assume that human translators preserve this formality across their translations. In doing so, we automatically create a dataset of over 1 million English sentences labeled for formality and show that classifiers trained on this data outperform those trained on manually labeled data.

Furthermore, we address issues present in the most popular source of manually labeled data for formality, namely that the dataset contains a significant topic bias. This means that classifiers trained on this data essentially take "shortcuts" and predict the formality of their input sentences based on the sentence's topic and not its style. We make the claim that formality is fundamentally defined by the *speaker relationship* and thus should not be affected by topic whatsoever.

We finish by demonstrating a proof of concept for a technique that would allow classifiers to ignore the topic of an input sentence and only see the style of a sentence. Finally, we suggest avenues for future research and outline guidelines for any future attempts at formality classification.

CHAPTER 2 : Formality Classification

2.1. English to Japanese Translation: The Formality Problem

Let's say that you're an English to Japanese translator. You have just been asked to translate a scene from a movie where an employee is talking to their boss. The line you are working on is one where the employee asks their boss "Have you seen the documents I sent you?". When translating this line from English into Japanese, it is not sufficient to simply translate this sentence's meaning alone. You must also include the information about the relationship between the speakers via the register of speech. If you mistakenly use the wrong register, the speech will come off as unnatural and rude.

Given your knowledge of the relationship between the two speakers, namely that the boss is of higher social status than the employee, you decide to translate this sentence with the Formal (Honorific) register form of the verb for "to see" and the Formal (Humble) register form of the verb for "to send".

私が	お送りした	書類を	ご覧になりましたか
I	Sent (Humble)	Document	Have you seen? (Honorific)

Now imagine you're tasked with the reverse problem. Imagine you are a Japanese to English translator and you are given the same exact sentence "私がお送りした書類をご覧になりましたか" and are asked to translate it into English. Given that the formal relationship between these speakers is explicitly encoded into this sentence, how do you properly reflect this in English?

2.2. Problem Definition: English Formality Classification

Unlike in Japanese, where there is an explicit distinction in register between informal and formal speech, English has no particular system for this distinction. Rather, the formality of a sentence is typically represented by a variety of stylistic factors, such as: proper spelling and grammar, proper punctuation and capitalization, lack of speech fillers such as "um" or "like", higher levels of politeness (although not always), use of precise, unambiguous speech,

and use of rarer words.

Informal	Formal
It's like, equally offensive to everyone!	I love all of them, and I can't name a single one.
i betta go and ask another stupid question lol..	I don't think he is in love with her.
It's cause ya got no sense.	I do not hate him but he makes me feel unwell.
Did you ever hear of a kleenix?	What exactly are you stating?
Did you REALLY pay money to see that?	I simply did not care enough to check.
pls answer i really wanna know.	Sadly, I no longer feel our unique connection.

Table 1: Examples of Informal and Formal sentences in English from Rao and Tetreault (2018)

However, while all of the above factors are typically *associated* with formality, they aren't necessarily indicative of it. For example, one can be very formal in their speech and still be impolite, one can speak formally while still using only common words, and it is clearly possible to be formal while still being ambiguous.

Additionally, a definition of formality that simply claims that these surface-level aspects comprehensively *define* formality is unsatisfying. A lack of contractions is a symptom, not a cause. In order to understand formality we need to understand the mechanisms working underneath that cause the emergence of these surface-level features.

2.3. What is Formality?: The "*speaker relationship*"

One thing that virtually all of the disparate definitions of formality have in common is that they all consist of elements of a relationship between speakers. The dimensions of power and solidarity (Brown and Gilman, 1960; Faruqui and Padó, 2012), the information and interpersonal (Biber, 1995), the desire to communicate unambiguously (Heylighen and Dewaele, 1999), and the amount of shared knowledge (Brown and Fraser, 1979) are all items that can be defined with respect to the relationship between the two speakers and not to the content of the speech.

Take, for example, a News broadcast. *Why* does the News anchor speak in a formal manner? Is it because the content of the News is formal? I propose that the speech here is formal

due to the relationship between the speaker and the listener. In this relationship there is a power dynamic, a lack of solidarity, a desire to communicate information unambiguously, and a lack of interpersonal intimacy – thus the sentence is spoken in a formal way. The formality of the sentence has nothing to do with the content of the news story itself. We know this to be clearly true, as we can recount such News stories to close family members accurately without the added degree of formality.

For the purposes of this work, I will be referring to this underlying circumscription of formality as the *speaker relationship*. I claim that the semantics of a sentence should be indicative of formality only in cases of *shallow formality* (Heylighen and Dewaele, 1999), a type of formality that is used as a ceremonial tool. For cases of *deep formality* I claim that the *speaker relationship* comprehensively defines the formality of an interaction. In defining the exact nature of this *speaker relationship* I take the hands-off approach common to modern computational linguists and rely on human annotations.

2.4. Sentence Formality Classification: Related Work

Early work on formality focused on the lexical level (Brooke et al., 2010; Brooke and Hirst, 2014). The lexical formality estimation problem is defined with respect to synonym pairs. That is, given two words with identical definitions, pick the more formal of the two.

To accomplish this task, Brooke et al. (2010) used lexicon induction, a technique whereby a lexicon of formal and informal terms is built by evaluating the average similarity of each term in the lexicon to a given set of formal and informal seed words and gradually building up that set of seed terms.

To calculate the lexical similarity of each term in the lexicon, Brooke et al. (2010) use cosine similarity of vectors returned by Latent Semantic Analysis (LSA) (Wolfe et al., 1998). LSA can be thought of as a Singular Value Decomposition (SVD) of the term-document matrix of a word. Performing LSA allows the cosine similarity score to focus on latent semantic variations across only those dimensions that vary the most across documents. In total the

Formality Score FS is calculated as:

$$FS'(w) = \sum_{s \in S, s \neq w} W_s \times FS(s) \times cos(\theta(w, s))$$

Where S is the set of all seed words, $\theta(w, s)$ is the angle between the two vectors w and s, $FS(s)$ is the formality score of the given seed word (formal seeds start with 1 and informal seeds start with -1), and W_s is the proportion of the total formality score of formal seeds vs. the total formality score of all seeds.

Brooke and Hirst (2014) used the average lexical formality score of all words in a sentence to estimate the formality of a sentence. While this method was shown to have respectable correlation with human annotations (Spearman $\rho = 0.49$), the repeated calculation of LSA was computationally expensive and there was demand for a simpler method.

This similar method was introduced by Pavlick and Nenkova (2015), who showed that lexical formality could be reliably predicted by analyzing the log odds ratio of the words appearing in known formal corpora such as Europarl (Koehn, 2005) versus known informal corpora such as Switchboard (Godfrey et al., 1992). The formula is:

$$FORMALITY(w) = log(\frac{P(w|REF)}{P(w|ALL)})$$

Where $P(w|REF)$ represents the probability of word w appearing in a given reference corpus REF. Similarly to Brooke and Hirst (2014), Pavlick and Nenkova (2015)'s lexical formality metric was also found to have respectable correlation with human annotations at the sentence level when using the same averaging technique (Spearman $\rho = 0.44$), but this was not an improvement accuracy over the previous state-of-the-art, rather an improvement in speed.

The first bona-fide attempt at full sentence formality classification was carried out by Pavlick

and Tetreault (2016). In order to answer the question of what features affected formality the most, Pavlick and Tetreault (2016) and Lahiri (2015) collected over 50,000 manual annotations for formality at the sentence level. The sentences given to annotators were sampled from four domains: News, Blog, Email, and Yahoo Answers. Each of the 10,000 sentences was graded by 5 separate annotators on a 7-point Likert Scale (-3 to 3) (Likert, 1932) and scores from all 5 annotators were averaged together to obtain a formality score for the given sentence.

Using this data, Pavlick and Tetreault (2016) conducted an empirical analysis of sentence level formality by training a ridge regression model on a combination of many different sentence-level features. This included F-Score (Heylighen and Dewaele, 2002), average log odds ratio (Pavlick and Nenkova, 2015), average word2vec embedding (Mikolov et al., 2013), Fleisch-Kincaid grade level (Kincaid et al., 1975), and many more. They aimed to empirically discern which of the previously mentioned features were most predictive for formality and to track how the predictive power of these features varied with changing domains. The results of their analysis are shown in Figure 1.

	Answers	Blogs	Email	News
ngram	0.84	0.85	0.84	0.91
word2vec	0.83	0.83	0.84	0.87
parse	0.70	0.89	0.74	0.89
readability	0.69	0.75	0.84	0.83
dependency	0.64	0.89	0.84	0.85
lexical	0.56	0.55	0.59	0.70
case	0.50	0.28	0.24	0.37
POS	0.49	0.74	0.67	0.74
punctuation	0.47	0.38	0.37	0.20
subjectivity	0.29	0.31	0.25	0.37
entity	0.14	0.63	0.34	0.72

Figure 1: Relative performance of each feature group across genres from Pavlick and Tetreault (2016). Numbers reflect the performance (Spearman ρ) of a ridge regression sentence formality classifier when using only the specified feature group, relative to the performance when using all feature groups.

They found that the most predictive features for linguistic formality across all domains

were ngrams features followed by average word2vec embeddings. This seemed to suggest that the **semantics** of a given sentence, not the style, are the most predictive of formality. However, given our definition of formality as being encapsulated by the *speaker relationship*, the predictive power of these semantic components is concerning.

In an attempt to address this weakness, Pavlick and Tetreault (2016) asked annotators to do 1,000 formal rewrites of informal sentences. They evaluated their classifier on these sentences and reported 88% accuracy for pairwise prediction, that is, given an informal sentence and its formal rewrite, determine which is the more formal of the two.

They claim that their 88% accuracy result shows that their ridge regression classifier uses stylistic not semantic elements for prediction. However, this claim is dubious, as the pairwise nature of the prediction task removes the possibility for systematic topical bias to affect the prediction outcome. In other words, since the semantic content of both the informal and formal sentences is identical, the 88% accuracy result only shows that the classifier *can* pick up on stylistic variations and not that it *does* use stylistic variations in its single-sentence classification.

Despite this shortcoming, the study conducted by Pavlick and Tetreault (2016) continues to serve a seminal role in formality estimation literature and is widely considered the state-of-the-art in sentence level formality estimation.

Now, with the relevant background out of the way, let us discuss the approach we will be taking in this work to improve sentence-level formality estimation.

CHAPTER 3 : Learning English Formality from Japanese

3.1. An Overview of The Approach

In order to project the information about the *speaker relationship* present in Japanese sentences onto their English translations we need three things. 1) We need a **Japanese-English Parallel Corpus** 2) We need a method of reliably performing **Japanese Formality Recognition** and 3) we need a **Sentence Formality Classification** algorithm that will learn a representation of English Formality given enough labeled data.

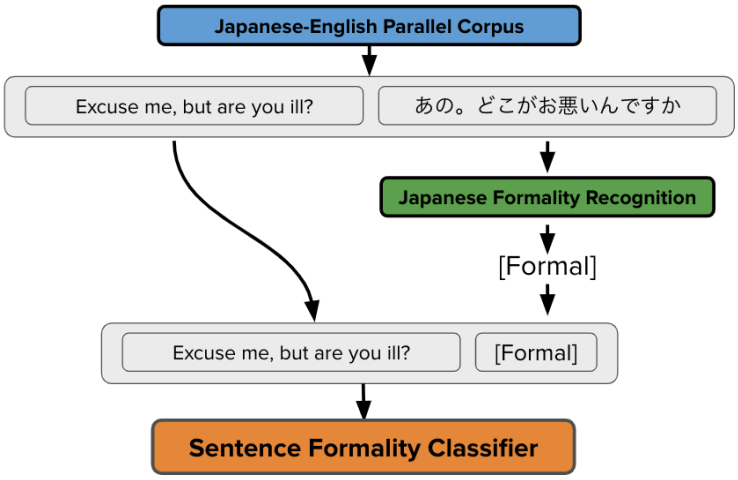

Figure 2: A general overview of our architecture

With these three items in hand we can use them as shown in Figure 2. The formality of the Japanese sentences can be identified and the labels can be attached to the English translation.[1]

We will start with the first of the three components necessary. The Japanese-English parallel corpus.

[1]Due to the limitations of binary classification and the fact that the Formal often circumscribes the Polite we consider both of these registers to be indicative of the "formal" class in this work. Future work should seek to incorporate both of these registers into classification as well as further investigate how our method could be used for politeness estimation

3.2. Step 1: Selecting a Japanese-English Parallel Corpus

In looking for a Japanese-English parallel corpus to use for our formality classifier, we would like a number of things. We would like one that:

1. Is large (preferably over 100k pairs)
2. Has good sentence alignments
3. Was constructed from reliable translations
4. Spans a variety of different topics and domains
5. Has many different *speaker relationships*

The fifth bullet point is particularly important here. For example, say we wanted to use the Kyoto Wiki Corpus to construct the formality dataset (Neubig, 2011). This corpus consists of over 500k manually translated sentences taken from Japanese Wikipedia articles about Kyoto. While this corpus has very reliable translations, good sentence alignments, and spans a variety of different topics, all of the sentences are from Wikipedia. Thus the only *speaker relationship* present in the corpus is that of a Wikipedia writer and Wikipedia reader. Training a formality classifier on this corpus would not give us any ability to parse out the *speaker relationship* of other types of sentences

This is also true for corpora like JParaCrawl (Morishita et al., 2019). JParaCrawl is the largest publicly available English-Japanese parallel corpus (with over 10 million sentence pairs). It was created by crawling the web and automatically aligning parallel sentences. While this corpus is incredibly large and spans a variety of different topics and domains, the filtration for parallel text ends up removing many of the sentences that would be the most interesting for our purposes. From manual inspection, it seems that a large percentage of this corpus consists of subtitles for pictures or legal descriptions. While these alignments may be reliable, this corpus also lacks the diversity of *speaker relationship* necessary to fully capture the stylistic variations that human translators induce when translating from

Japanese to English.

That leaves us with only one option, the Japanese English Subtitle Corpus (JESC) (Pryzant et al., 2018). JESC was created by scraping data from several repositories, where amateur fan translators can freely upload their own translations for *anime*, *manga*, and television programs.[2]

There are many reasons why JESC is very well suited for our study. To start with, the corpus is the second largest publicly available Japanese-English parallel corpus (3.2M sentences) only second to the previously mentioned JParaCrawl (Morishita et al., 2019). Secondly, the translations span a wide array of domains due to the wide variety of media present on the subtitle sites. This allows the corpus to not be dominated by any one topic and lets classifiers trained on this data focus more on stylistic elements. Finally, and most importantly, due to the large number of characters in a given book or television show, there are many different *speaker relationships* represented in this corpus. This allows our classifiers to have the maximum possibility of learning the ways in which human translators interpret each of these *speaker relationships*.

There are two shortcomings to using JESC, both of which do not outweigh the positives. To start with, the alignments of the sentences are of questionable quality. This is a natural consequence of sampling parallel sentences from subtitles; Sometimes the ordering of information is changed to improve the fluency of speech (this is especially common in Japanese). While these misalignments may be an issue for tasks such as machine translation, for our purposes, this is not an issue.

Remember that formality is defined by a given *speaker relationship* and not by semantic content. Thus, as long as the misalignment is localized to a given interaction between two characters, the formality levels of the misaligned sentences will still be accurately aligned, even if the meaning of the sentences are completely different.

[2]These include kitsunekko.net, d-addicts.com, opensubtitles.com, and subscene.com

The second major shortcoming is that JESC consists of subtitles submitted by amateur translators. Unlike the previous shortcoming, this one is definitely a concern, as amateur translators may not completely understand the cultural context of certain interactions and tend to take more liberties in their translations. While this is a very serious issue with using JESC for these purposes, we do not believe it is enough to warrant it being thrown away.

That being said, it is worth keeping in mind that the method and results reported in this work are achieved using a corpus of amateur translations. JESC will serve as a lower bound for the possibilities of future work with Japanese-English Parallel Corpora.

Now that we have finished selecting our parallel corpus, we must turn our focus to the second component, the Japanese Formality Recognition algorithm.

3.3. Step 2: Recognizing Japanese Formality

Formality in Japanese can be broken up into four distinct registers:

- **Informal**: used primarily with close friends, family, and those who are younger.
- **Polite**: 丁寧語 (*teineigo*) used with acquaintances and those of generally equal social status.
- **Formal (Honorific)**: 敬語 (*keigo*) used by persons of low social rank in talking to a superior *about a superior's actions* (Bloch, 1946).
- **Formal (Humble)**: 謙譲語 (*kenjougo*) used by persons of low social rank in talking to a superior *about their own actions* (Bloch, 1946).

As we can see from this taxonomy, **the register of a sentence in Japanese communicates fine-grained information about the *speaker relationship***. If the speaker is of higher status than the listener, the speaker will use the Informal form. If the speakers are of equal status and not intimate, they will use the Polite form. These principles are

unaffected by the topic save for very specific and rare scenarios.[3]

In order to identify the register of a sentence in Japanese, typically we look to the conjugation of the sentence ending verb. This sentence ending verb is required for a grammatically correct Japanese sentence, and thus serves as a reliable marker for sentence formality.

Informal	**Polite**	**Formal (Honorific)**	**Formal (Humble)**
探す	探します	お探しになります	お探しします
sagasu	*sagashimasu*	*o-sagashi ni narimasu*	*o-sagashi shimasu*

Table 2: Conjugations of the Japanese verb "To Search" (探す) in the Present tense for Informal, Polite, and Formal registers

While the conjugations of the Polite register are common and generally easy to spot, the Formal register is a bit more difficult. While the Formal register *may* involve periphrastic constructions, or suffixes that are identical to those used in passive and causative verb forms (Prideaux, 2017), the most typical formulation of the Formal register, by far, is in alternate choices of verbs. Some examples are given below:

Informal	**Polite**	**Formal (Honorific)**	**Formal (Humble)**
見る	見ます	ご覧になります	拝見します
miru	*mimasu*	*go-ran ni narimasu*	*haiken shimasu*
言う	言います	おっしゃいます	申します
iu	*iimasu*	*osshaimasu*	*moushimasu*

Table 3: Conjugations of the Japanese verb "To See" (見る) and "To Say" (言う) in the Present tense for Informal, Polite, and Formal registers. Note that the verb stems used in the Formal register differ from those used in Plain and Polite

While an exhaustive list of alternative Formal verbs and other Formal constructions is beyond the scope of this paper, curious readers can find a sizeable list in the 日本語表現文典 *nihongo hyougen bunten* ("Dictionary of Japanese Phrases") (Okamoto, 1944). While many of the constructions found in this dictionary are very rare and are becoming ever more so, it is still the most comprehensive reference for Japanese formality.

Previous work done by Feely et al. (2019) proposes to recognize formality via a simple

[3]Most of these scenarios would be considered examples of *shallow formality* (Heylighen and Dewaele, 1999)

	Informal	Polite	Formal
Precision	1.00	0.82	0.72
Recall	0.74	0.91	0.97
F1	0.85	0.86	0.83

Table 4: Evaluation scores of labels produced by Feely et al. (2019)'s formality recognizer (as reported by Feely et al. (2019)) compared to gold test set labels for each register.

substring search on Japanese text tokenized with MeCab (Kudo, 2006) for a set of Formal, Polite, and Informal key verbs and suffixes. Given an input sentence in Japanese, if a Formal register key is present, the sentence is labeled Formal. Otherwise, if a Polite register key is present it is labeled as Polite, and if an Informal register key is present it is labeled as Informal. If no key is present, then no label is given. The string matching keys used in their Japanese formality parser are listed in Table 5.

The authors of this paper hired a Japanese linguist to annotate a set of 150 sentences taken from the Tanaka Corpus (Tanaka, 2001). The self-reported accuracy of their recognizer in predicting these manual annotations is listed in 4.

Formality	Verb Forms
Informal	だ, だった, じゃない, じゃなかった, だろう, da, datta, janai, janakatta, darou だから, だけど, だって, だっけ, そうだ, ようだ dakara, dakedo, datte, dakke, souda, youda
Polite	です, でした, ない, なかった, ます, ました, ません, desu, deshita, nai, nakatta, masu, mashita, masen ましょう, でしょう, ください, なさい, である, ではない mashou, deshou, kudasai, nasai, dearu, dewanai
Formal	ございます,いらっしゃいます,おります,なさいます,致します gozaimasu, irasshaimasu, orimasu, nasaimasu, itashimasu ご覧になります,拝見します,お目に掛かります, goran ni narimasu, haiken shimasu, o me ni kakarimasu おいでになります,伺います,参ります,存知します,存じ上げます, oide ni narimasu, ukagaimasu, mairimasu, zonji shimasu, zonji agemasu 召し上がります,頂きます,頂く,頂いて, meshi agarimasu, itadakimasu, itadaku, itadaite, 差しあげます,下さいます,おっしゃいます,申し上げます sashi agemasu, kudasaimasu, osshaimasu, moushi agemasu

Table 5: Verbs and verb suffixes used by Feely et al. (2019) to recognize Japanese sentences of specific registers

While the size of the evaluation dataset used by Feely et al. (2019) is concerningly small, it does seem that string searching for their set of keys performs well on Japanese Formality recognition. However, their set of keys is somewhat clunky. Many keys overlap with each other (both within and across registers), some keys tend to get broken up by MeCab (and thus are never found in tokenized text), some keys are too short and thus are prone to false positives. It is clear from this set of keys that we can further streamline this method.

In order to streamline this we must keep in mind our task. We are interested in the set of keys that most consistently influence translators to write their English translations with formal language. In other words, our set of string matching keys should exclude all Japanese verb endings that are not predictive of English formality.

In order to determine which of the proposed keys from the Feely et al. (2019) set are predictive of English formality, we fine-tune a BERT classifier[4] on a set of manually labeled English formality data from Pavlick and Tetreault (2016) and Lahiri (2015) and use it to label the English half of JESC. We then train a Logistic Regression classifier with character ngrams features on the Japanese sentences to predict these labels.[5] We sort the features by weight and rank each key from Feely et al. (2019) by how predictive it was for formality in English. A diagram of this structure is in Figure 3 and results are listed in Table 6.

We can see from this result that many keys in the recognizer from Feely et al. (2019) do not correspond well to English formality. In particular, the keys that denote the Informal register are wildly inconsistent and the keys that denote the Formal register are too rare to even pass our frequency threshold. If we were to use this key set as our Japanese Formality Recognizer we would be assigning many formal English sentences an informal label and barely any sentences would be labeled as formal. Due to this, we decided to restrict our set to only consist of Polite register keys and to label all sentences that contain one of those keys as formal and all sentences that do not contain a key as Informal. The Polite register

[4]Refer to Section 3.5 for more information on BERT and why we use it

[5]N-grams were thresholded to at least 500 appearances in 200k sentences. Scikit-learn (Pedregosa et al., 2011) was used to train a logistic regression, using the Stochastic Average Gradient solver.

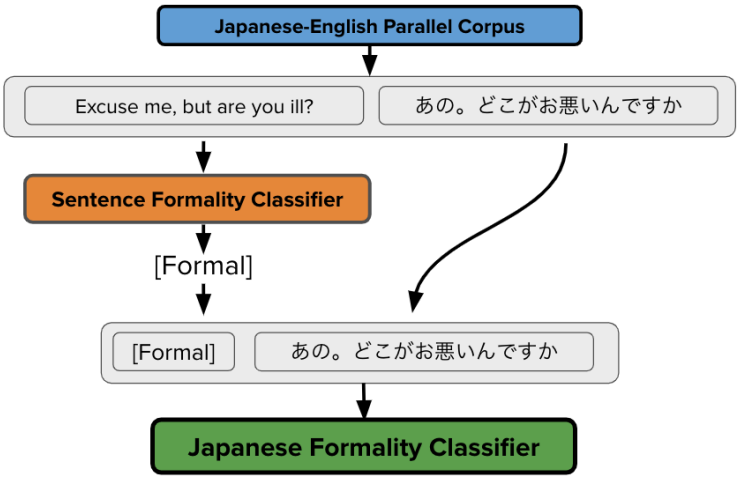

Figure 3: The architecture for our Empirical Formality Key Derivation Experiment

keys that were selected for use in our recognizer are marked by asterisks in Table 6 and are listed in full in Table 7.

Note that since most Formal verb substitutions take Polite register endings, this recognizer can successfully detect both Polite and Formal register sentences by only using these seven keys. Additionally, the absence of these keys does, in fact, indicate informality, since the sentence ending verb is required for all complete sentences. In other words, with just these seven keys, we have a very powerful recognizer. In addition, since many of our keys are unique to verb endings and rarely appear within or across words, our recognizer can operate on raw untokenized text, making it incredibly fast and scalable.

3.4. Step 3: Classifying Sentence Formality

Now that we have both a reliable way of recognizing Japanese formality and a high quality Japanese-English parallel corpus, all we need left is an algorithm that will properly learn a representation of sentence formality given enough labeled data.

The architecture we will be using for this is the popular BERT architecture (Devlin et al., 2019). There are a few reasons why we will be using BERT. First, it is a pre-trained model. This means that we do not need much data in order for BERT to learn our task, as it

Key	Rank	Percentile	Register
だろう	563	93	*Informal*
だけど	1257	86	*Informal*
ません*	1991	77	*Polite*
でしょう*	2258	74	*Polite*
ました*	2995	66	*Polite*
です*	3093	65	*Polite*
ます*	3273	63	*Polite*
だって	3408	61	*Informal*
でした*	3464	61	*Polite*
ください	3481	60	*Polite*
だった	3837	56	*Informal*
だから	4106	53	*Informal*
そうだ	4121	53	*Informal*
ましょう*	4153	53	*Polite*
である	4186	52	*Polite*
ようだ	4908	44	*Informal*
なさい	5371	39	*Polite*
ない	5594	36	*Polite*
なかった	5767	34	*Polite*
だ	7492	15	*Informal*
じゃない	7866	11	*Informal*

Table 6: Regular expression keys from the recognizer proposed by Feely et al. (2019) that appeared more than 500 times in 200k sentences ranked by how predictive they are for English formality (out of 8801 character sequences). Asterisks denote the seven keys we selected for use in our recognizer.

です	でした	ます	ました	ません	ましょう	でしょう
desu	*deshita*	*masu*	*mashita*	*masen*	*mashou*	*deshou*

Table 7: The complete list of the 7 keys used in the final Japanese Formality Recognizer

already comes initialized with a basic understanding of English. The second reason is that it uses contextual representations of words rather than fixed representations. This makes it much more powerful than both ridge regression (Pavlick and Tetreault, 2016) (which uses only fixed vector space models) and lexical methods such as lexicon induction and log-odds ratio (Brooke et al., 2010; Pavlick and Nenkova, 2015) (neither of which take the ordering of the words in a sentence into account).

3.5. Creating and Evaluating the "Japanese Formality Corpus"

Now that we have the three components in place (Corpus, Recognizer, and Classifier) we can begin to put the pieces together and test the efficacy of our method.

After running our Japanese formality recognizer on the Japanese-English Subtitle Corpus, we were able to label 520K sentences as formal and 2.28M sentences as informal. We project these binary labels onto the English sentences, discard the Japanese, and down-sample to create an equal ratio of formal to informal sentences, leaving 1.04M sentences. This is the largest labeled dataset for formality ever constructed.

We will refer to this dataset throughout the rest of the work as the JFC.

In order to evaluate the accuracy of the labels in the JFC we will fine-tune a BERT classifier on JFC and compare the output of that classifier to other BERT classifiers fine-tuned on other popular datasets. If the performance achieved by the BERT classifier fine-tuned on JFC is better, then our dataset is of high quality and our approach has been validated.

3.6. Data Comparisons

To evaluate our approach we compare the JFC to three other formality datasets from previous literature.

Our first comparison dataset is the GYAFC (Rao and Tetreault, 2018). This dataset consists of 110K formal-informal sentence pairs and has seen significant use in recent work on formality style transfer. (Xu et al., 2019; Niu et al., 2018; Wang et al., 2019; Luo et al., 2019; Cheng et al., 2020) It was created by asking annotators to make formal rewrites of informal sentences from Yahoo Answers. For the purposes of binary classification we label all formal rewrites as formal and all informal sentences as informal. This dataset is used for both fine-tuning (Table 8) and evaluation (Table 9).

Our second comparison is the dataset compiled by Faruqui and Padó (2012) which uses a similar methodology to our approach, but with another language pair. This dataset

consists of 500K parallel English-German sentences sampled from 110 novels available from Project Gutenberg (English) and Project Gutenberg-DE (German). Faruqui and Padó (2012) marked these sentences as formal or informal based on the corresponding German formal/informal pronouns *sie* and *du*. We exclude all sentences that lack one of these German pronouns and therefore end up with 34K informal and 55K formal. We then down-sample to an equal ratio, leaving 67K sentences. This dataset is used for fine-tuning only (Table 8 and 9)

Our third comparison is the manually labeled dataset from Lahiri (2015) and Pavlick and Tetreault (2016). In this dataset, annotators give a 7-point Likert scale (Likert, 1932) judgement on the formality of a given sentence with five annotators per sentence. Sentences are drawn from four different domains (1821 Blog, 2775 News, 1701 Email, and 4977 Yahoo Answers). For the purposes of our experiment the average of these five judgements is taken and thresholded at 0 to obtain a binary class label. This dataset is used both for fine-tuning (Table 9) and for evaluation (Table 8).

3.7. Results

For all experiments we use the "bert-base-uncased" model via the HuggingFace transformers library[6] (Wolf et al., 2019). We fine-tune for 2 epochs with a learning rate of $5 * 10^{-5}$.

Method	**Blog**	**Email**	**News**	**Yahoo**	**Total**
Fine-Tuned on GYAFC	71.81	77.55	85.91	54.72	72.88
Fine-Tuned on German-English	65.39	76.92	80.40	43.01	66.34
Fine-Tuned on Japanese-English (Ours)	**73.18**	**80.00**	**86.54**	**54.76**	**73.98**

Table 8: F1-Scores for BERT classifiers fine-tuned on different datasets. We evaluate on the manually labeled data from Pavlick and Tetreault (2016), which is binned by domain. For the full results please see Table 12 in the Appendix

In Table 8 we show the results of the fine-tuned BERT classifiers when evaluated on the manually labeled data. We see that our method significantly outperforms the use of a German-English parallel corpus and performs comparatively well to fine-tuning on GYAFC.

[6]https://github.com/huggingface/transformers

Method	Precision	Recall	F1-Score
Fine-Tuned on Manually labeled data	**95.07**	18.73	31.29
Fine-Tuned on German-English	56.14	46.39	50.80
Fine-Tuned on Japanese-English (Ours)	64.17	**57.44**	**60.62**

Table 9: Precision, Recall, F1-Score for BERT classifiers fine-tuned on different datasets evaluated on GYAFC data (Rao and Tetreault, 2018)

We hypothesize that the low performance of the German-English parallel corpus is due to differences in the formal and informal markers. While Japanese formality is indicated by verb suffixes, German formality is indicated by formations of the second person pronoun "you" and thus occurs only in direct address sentences. This introduces a bias in the distribution of sentences in the training set, as only sentences which contain this marker can be reliably labeled. By contrast, all complete Japanese sentences contain a verb and therefore classifiers trained on Japanese-English parallel data do not contain this bias.

In Table 9 we show the results of our classifiers when evaluated on GYAFC. We see that our Japanese-English data outperforms the Manually Labeled data in both Recall and F1 Score. We hypothesize that this is due to a topic bias in the Manually Labeled data. GYAFC is a corpus constructed from Yahoo Answers and, since the majority of the Yahoo Answers subset of manually labeled data is informal, the classifier trained on Manually Labeled data likely takes Yahoo Answers-related subject matter as being predictive of informality. In the next chapter we further investigate this topic bias and look at the degree to which it can be explicitly controlled through the use of adversarial decomposition.

3.8. Takeaways

In this chapter we showed that it was possible to train a formality classifier in English without any access to manually annotated data. We also showed that such a classifier generalizes well across domains, outperforming datasets from previous literature in all four domains of manually labeled data.

We explain that our technique is fast and scalable and that it works even when the corpus being used has frequent small misalignments and questionable translation quality.

Finally, we claim that this technique can be applied in any language that has Japanese parallel sentences, setting the stage for the development of massively multilingual formality classifiers in the future.

CHAPTER 4 : Separating Formality from Topic

4.1. Topic Bias in Human Annotations

To investigate why manually labeled formality data from Pavlick and Tetreault (2016) and Lahiri (2015) did so poorly when evaluating on GYAFC (Rao and Tetreault, 2018) we borrow the same experimental setup from our empirical derivation of formality keys (See Figure 3). However, instead of querying for the rank of a given set of formality keys, we will sort every character ngram in Japanese and look at the top 5 most predictive 4-grams 3-grams and 2-grams for formality. The top five most predictive character ngrams in each of these categories are listed in Table 10.

4-grams	**3-grams**	**2-grams**
されてる *have been*	2時間 *two hours*	今朝 *this morning*
見つかっ *found (informal)*	い眠り *sleep*	捜査 *investigation*
アメリカ *America*	可能な *possible (adj.)*	作戦 *tactics, strategy*
における *as for, regarding, in*	大学の *of the university*	政府 *government*
殺された *was killed*	以外に *with the exception of*	事故 *accident*

Table 10: The Japanese character sequences that were most predictive of English formality (as understood by manually labeled data). We see that topic-oriented words appear here instead of the Polite or Formal verb endings.

If the classifiers trained on manually labeled data were truly picking up on the underlying *speaker relationship* we would expect to see the Polite verb suffixes and the Formal register verbs represented in this list. However, from the results of this experiment, we can clearly see that classifiers fine-tuned on the manually labeled data in English do not pick up on this stylistic component, instead listing News-like words like “America”, “investigation”, and “was killed” as being highly predictive of formality in Japanese while the verb suffixes rank much lower (See Table 6). This confirms our suspicion of a topical bias, as News was the domain that had the highest average formality score. This also supports our observations from the previous chapter when we noticed that the most predictive features for formality reported by Pavlick and Tetreault (2016) were ngrams and word2vec embeddings, both of which are semantic and not stylistic features.

This raises serious concerns with regards to the use of this set of manually labeled formality data as a benchmark for evaluation and further underscores the necessity of semi-supervised sources such as GYAFC for comprehensive evaluation of the accuracy of formality classifiers.

It is clearly not a good idea for a formality classifier to focus on topic-oriented vocabulary, as is evidenced by the manually labeled data classifiers' inability to generalize. That being said, it is **impossible** to prevent formality classifiers from picking up these topic-oriented correlations. If there is a topic bias in a training dataset, classifiers trained on it will contain that topic bias – and there will always be *some* topic or semantic bias in any set of training data. Even if the number of formal and informal sentences is perfectly equalized across every single sub-domain, it is always possible to find a new semantic dimension to group the sentences across which a dataset is biased.

The only way to truly move forward against this problem of topic bias in formality classification is by finding a way to show classifiers only the style of a sentence, without any of the semantic aspects.

4.2. Adversarial Decomposition: The Problem Statment

At a high level, adversarial decomposition seeks to split up the vector representation of a given sentence into two latent vector representations. One of these vectors denotes the form of the sentence, we will call this the "style vector", and the other denotes the meaning of the sentence, we will call this the "meaning vector".

To borrow the formal statement from Romanov et al. (2018), let X^a be a corpus of sentences $x_i^a \in X^a$ in Formal English $f^a \in \mathcal{F}$, and X^b be a corpus of sentences $x_i^b \in X^b$ in Informal English $f^b \in \mathcal{F}$. We assume that the sentences in both X^a and X^b have the same distribution of meaning $m \in \mathcal{M}$. The form f, however is different and generated from a mixture of two distributions:

$$f_i = \alpha_i^a p(f^a) + \alpha_i^b p(f^b)$$

Where f^a and f^b are the different forms of English (in our case, formal and informal English). We say that a sample x_i has form f^a if $\alpha_i^a > \alpha_i^b$ and that it has form f^b if $\alpha_i^b > \alpha_i^a$

Thus the goal of dissociating meaning and form is to learn two encoders $E_m : \mathcal{X} \to \mathcal{M}$ and $E_f : \mathcal{X} \to \mathcal{F}$ for the meaning and form correspondingly, and a generator $G : \mathcal{M}, \mathcal{F} \to \mathcal{X}$ such that

$$\forall j \in a, b, \forall k \in a, b : G(E_m(x^k), E_f(x^j)) \to \mathcal{X}^j$$

4.3. Adversarial Decomposition: The ADNet Architecture

To perform adversarial decomposition, we use the ADNet architecture[1] (Romanov et al., 2018). The ADNet architecture is based on adversarial-motivational training, GAN architecture (Goodfellow et al., 2014), and adversarial autoencoders (Makhzani et al., 2015).

There are four main components in this architecture. The Encoder E is a Gated Recurrent Unit (GRU) (Chung et al., 2014) model that produces a single hidden vector h. This vector h is passed through two different fully connected layers to get the form vector f and meaning vector m. The Generator G is also a GRU unit and it attempts to reconstruct the original hidden vector h from the input f and m to ensure that the Encoder properly encodes the content of sentence in these vectors.

The Discriminator D is given only the meaning vector m as input and is tasked with predicting the form vector f given m. The negative loss of this discriminator is passed back into the Encoder so that it can better optimize to create meaning vectors that are maximally distant from the form vectors. In addition to this, the Motivator M is given only the form vector f and uses it to classify the form of the sentence and the positive loss is passed back into the Encoder. Thus while the Discriminator influences the Encoder to not include form information in the meaning vector, the Motivator actively encourages the Encoder to encode form information in the form vector.

[1] https://github.com/text-machine-lab/adversarial_decomposition

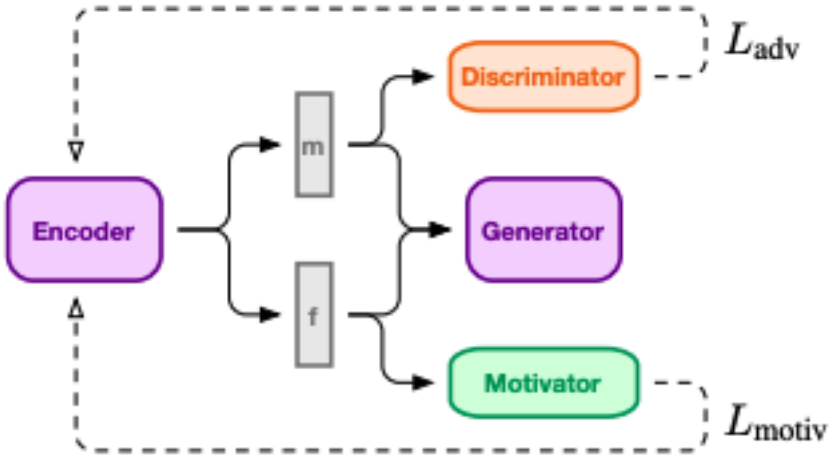

Figure 4: A Diagram of the ADNet architecture from Romanov et al. (2018)

4.4. Experiments

We train ADNet on three datasets: Manually labeled data (Pavlick and Tetreault, 2016; Lahiri, 2015), GYAFC (Rao and Tetreault, 2018), and on our JFC. A test set of size 10,000 is held out for JFC and GYAFC and a proportionally sized 1,000 sentence test set is held out for manually labeled data.

We train for 500 epochs with a learning rate of 0.001 and Dropout parameter of 0.2 (Srivastava et al., 2014). The size of both the style and meaning embeddings are set to 128 units each.

Before training starts, we extract the initial vector representations of every sentence in the test set. These are referred to as the "untrained" vectors. Then, once our model is finished training, we use the trained model to extract the "style" and "meaning" vectors for every sentence in the test set. This is repeated for all three datasets.

4.5. Results

To get a good understanding of what the adversarial training was able to capture we ran a Principal Component Analysis (Wold et al., 1987) on all vectors to lower the dimensionality

to 50 and then used t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008) to visualize the results of the training for all three vector classes. The results of this visualization are shown in Figure 5

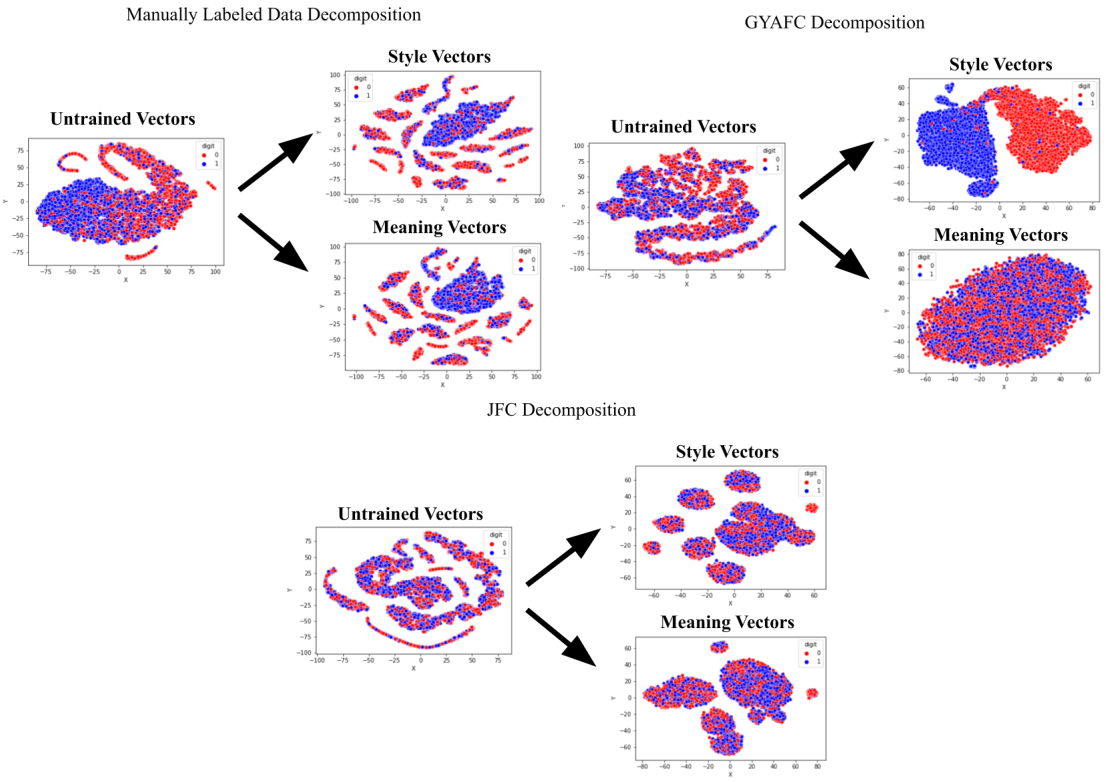

Figure 5: t-SNE plots for the three datasets of interest (Manually labeled, GYAFC, and JFC) from before and after training with ADNet (Romanov et al., 2018) A red dot indicates an informal sentence, a blue dot indicates a formal sentence.

From these plots we can clearly see that the stylistic dimension of the sentences in GYAFC is very pronounced and is clearly separable. This is not surprising, as GYAFC is composed entirely of pairs of sentences that have identical meaning but only vary across their style.

Our Japanese Formality Corpus had a more interesting result. While not as clearly separable as GYAFC, there was a significant change between the untrained vectors and the trained vectors. In particular, the trained vectors seem to form cohesive clusters while the untrained vectors are much more haphazardly distributed.

The degree to which the plots change in JFC does show that the ADNet model was somewhat able to separate topic from style to some degree. However, it is clear that JFC still holds plenty of implicit topical biases that were not completely separated out from the corpus.

After this experiment, we used the three vector types as inputs to a Logistic Regression classifier and evaluated their prediction accuracy on their own held out test sets. The results are listed in Table 11.

Dataset	Untrained	Style	Meaning	Gain from ADNet Training
Manual	**69.69**	69.96	70.54	+0.27
JFC	56.72	63.33	63.32	+6.61
GYAFC	62.55	**96.38**	**95.86**	**+33.83**

Table 11: Accuracy of a Logistic Regression classifier when using each type of latent vector to predict binary formality. Accuracy is evaluated on a held out test set of each dataset

The first thing to note about these results is that the meaning vectors consistently do about as well as the style vectors in the prediction task.

This is likely due to a quirk in the architecture. Remembering back to Figure 4, the Discriminator D module explicitly optimizes the meaning vector m to not be able to predict the form vector f. However, the Motivator M only motivates f to be able to predict the label and ***does not*** **punish** m **if it can predict the label**. This means that m likely is being trained to be the exact polar opposite of f. This is a pattern that could easily be picked up on by a Logistic Regression Model.

At the end of the day, the above explanation is only a hunch. Future work should investigate ways to properly guarantee that only semantic information is present in m and only stylistic information is present in f.

Once again returning to Table 11, we can judge the amount of topical bias present in a dataset by the extent to which the accuracy is improved after ADNet training.

We can see that the manually labeled data did not get any more accurate at the classification task, as the model was unable to learn anything due to the heavy correlation between the

semantic elements of the sentences and their formality labels.

On the opposite end of the spectrum, GYAFC had exceedingly high performance gains after ADNet training, reaching an astounding 96.38% accuracy for non-pairwise binary formality classification. This is the highest accuracy for binary formality classification I have seen on any dataset across all relevant literature. A result like this bodes well for the future usage of the technique.

Finally, in-between the two extremes is the JFC, which gained a modest 6% in performance. This implies that JFC was transformed a small amount by the model but not to the extent of GYAFC. Once again, this is expected. It is unrealistic to assume that an automatically generated corpus like JFC would stack up to a monolingual parallel corpus like GYAFC which is stylistically separable *by construction*.

4.6. Takeaways

In this chapter we attempted a proof of concept study into developing a topic-agnostic formality classifier. While the method was only fully successful in separating out form in a corpus specifically designed for this task, we still believe our results with that dataset bode well for future research in this area.

Without a method that explicitly separates the meaning from the style of a sentence, it is impossible to guarantee that a classifier is not topically biased in some way, regardless of if the labels correspond to the underlying *speaker representation* or not. If there exists any correlation between topic and label, a classifier will learn it and will use it in some way to educate its predictions.

In order to truly approach a classifier that evaluates formality in a way that generalizes, it is crucially necessary to minimize this topical bias.

CHAPTER 5 : Conclusion

Through the use of Japanese-English Parallel Corpora, we constructed the Japanese Formality Corpus (JFC), the largest formality corpus of its kind, consisting of over one million labeled sentences. We show that classifiers trained on this dataset outperform classifiers trained on other datasets from the relevant literature.

In addition, through our discussion of the *speaker relationship*, we demonstrate the shortcomings of the current collection of manual formality annotations. We suggest that manual annotations without proper context may not be the proper method for developing robust stylistically-focused sentence formality classifiers. We show that there exists a middle ground between topically biased manually labeled formality datasets and perfectly unbiased but expensive monolingual parallel corpora.

Finally, we demonstrate a proof of concept for the use of adversarial decomposition to decouple topic from style, allowing formality classifiers to condition only on style and ignore spurious topic correlations. We show a path forward for future work on topic-agnostic classifiers with the hope that the formality classifiers of the future will live up to this potential.

Sentence formality classification is a difficult task. Sometimes the formality level of a given sentence truly is ambiguous. Without access to the broader context of a given interaction, there is not much a classifier can do from looking at stylistic components alone, especially when those components are scarce or absent altogether. That being said, formality classification research is *far* from hitting that performance ceiling.

There are many ways forward for future research to improve sentence formality classification. Future work can and should:

- Use the technique described in this work on other honorific languages such as Korean

or Javanese[1].

- Construct a more fine-grained Japanese Formality recognizer and use it to build a Categorical formality dataset
- Attempt to apply the JFC to Politeness Classification
- Investigate more into how Adversarial Decomposition can help sentence formality classifiers generalize
- Investigate how Machine Translation performance and style improves with reliable formality classification applied as a preprocessing step
- Investigate if there are topical biases present in current formality style transfer models

On top of this, even more clever ideas are coming out of the literature on formality classification, such as Online Target Inference (Niu and Carpuat, 2019) and Cross-Cultural Transfer Learning (Ringel et al., 2019).

If improvements such as these are made to the accuracy and robustness of sentence formality classification, it will have a broad impact on a variety of tasks. We will see better dialogue agents, better grammar correction, better formality style transfer, better machine translation, better information retrieval, and much more. Significant potential for these advancements exists; I hope this thesis has helped to lay their groundwork.

[1]The language of the people of the island of Java (https://en.wikipedia.org/wiki/Java)

APPENDIX

The full results for the different classifiers evaluated on Manually Labeled data are reported below in Table 12

Data	Blog	Email	News	Yahoo	Total
Baselines					
All Formal	70.38	75.42	85.43	41.81	65.25
All Informal	62.73	56.59	40.55	84.77	68.05
Manually Labeled Data					
Logistic Regression (SOTA)	78.56	80.29	88.26	49.54	77.27
Feed-Forward Neural Network	76.98	82.11	87.95	59.83	78.14
BERT	**81.02**	83.52	89.12	65.32	**81.82**
RoBERTa	78.74	83.26	89.01	**67.68**	81.26
GYAFC					
Logistic Regression	68.81	77.53	85.69	52.25	70.97
Feed-Forward Neural Network	65.50	73.32	82.92	53.54	69.80
BERT	71.81	77.55	85.91	54.72	72.88
German-English					
Logistic Regression	71.85	76.21	84.07	43.72	67.49
Feed-Forward Neural Network	72.48	75.53	82.62	44.27	67.43
BERT	65.39	76.92	80.40	43.01	66.34
Japanese-English (Ours)					
Logistic Regression	73.59	79.27	86.31	51.25	71.79
Feed-Forward Neural Network	67.22	72.31	80.57	49.02	66.52
BERT	73.18	80.00	86.54	54.76	73.98
Monolingual Pre-Training					
Logistic Regression	68.81	77.30	85.54	51.66	70.71
Feed-Forward Neural Network	76.39	82.63	88.18	62.66	78.74
BERT	79.25	84.30	88.73	66.11	81.39
German-English Pre-Training					
Logistic Regression	76.36	78.30	86.62	53.27	74.82
Feed-Forward Neural Network	76.54	81.12	87.80	61.51	78.04
BERT	80.00	82.77	88.94	63.18	81.01
Japanese-English Pre-Training					
Logistic Regression	74.86	82.22	86.64	52.53	73.66
Feed-Forward Neural Network	75.52	81.67	88.23	59.36	77.36
BERT	78.37	**85.59**	**89.54**	62.47	81.20

Table 12: F1-Score of our binary sentence level formality classifiers on the four domain areas (Blog, Email, News, and Answers) of a held out test set of Manually Labeled Data. "Pre-Training" models were first fine-tuned on the specified dataset then fine-tuned on the training set of Manually Labeled Data

GLOSSARY

ADNet Adversarial Decomposition Net. x, 24–28

adversarial a class of machine learning modules which optimize to lower accuracy on a certain undesirable task while simultaneously optimizing to increase accuracy on a desirable task. v, 20, 23–25, 29, 30

agnostic in machine learning, not concerned with or ambivalent to (e.g. topic-agnostic). v

alignment in a parallel corpus, the degree to which two parallel sentences contain identical semantic information. 10, 11, 20

BERT Bidirectional Encoder Representations from Transformers. viii, ix, 16, 18–20

bias a systematic distortion of a statistical result. In machine learning, a reliance on features that are not truly predictive of the output label due to false correlations present in training data. v, 2, 8, 20, 23, 27–29

classifier a machine that automatically infers the class of an input item. v, viii, ix, 1, 2, 8, 10, 11, 15, 18–23, 27–31

conjugation the variation of the form of a verb in an inflected language by which are identified the voice, mood, tense, number, and person. viii, 13

cosine similarity the similarity between two vectors as defined by the cosine of the angle between them. 5

domain a broad term used to encapsulate a certain stylistic or semantic commonality between sentences (i.e. Medical Domain, Scientific Domain). viii, ix, 7, 10, 11, 19, 20, 23, 31

embedding a vector representation of an input item (e.g. word, sentence). 7, 8, 25

generalize in machine learning, to perform well on new, unseen data. 20, 23, 28, 30

GYAFC Grammarly Yahoo Answers Formality Corpus. ix, x, 18–20, 22, 23, 25, 26, 28

hidden in machine learning, a vector representation of an input that is internal to a given architecture. 24

JESC Japanese English Subtitle Corpus. 11, 12, 15

JFC Japanese Formality Corpus. x, 18, 25–30

key the target sequence of words or characters to be used in a string search. viii, x, 14–17, 22

latent (of a quality or state) existing but not yet developed, the underlying semantic or stylistic components of a word or sentence. ix, 5, 23, 27

learning rate in machine learning, the amount that the weights of a network are updated during training. 19, 25

lexical relating to the words or vocabulary of a language. 1, 5, 6, 17

lexicon the complete set of meaningful units in a language. 5, 17

LSA Latent Semantic Analysis. 5, 6

ngram a contiguous sequence of n items from a given sample of text or speech (typically characters or words). 8, 15, 22

parallel corpus A large collection of sentences that are related to one another by meaning. These can differ in language or style.. v, 2, 9–12, 16, 19, 20, 28

recognizer a machine that automatically detects surface-level features and uses them to recognize the class of an input item. This differs from a classifier in that the class of the input item for a recognizer is well defined and unambiguous. viii, 15–18, 30

register a variety of a language or a level of usage, as determined by degree of formality. viii, 3, 9, 12–16, 22

semantic relating to meaning in language or logic. 5, 8, 22, 23, 27, 28

semi-supervised data data which are assigned labels automatically using information from a supervised set of data. 2

substring search to look for a key sequence of words or characters in a large amount of text. 14

suffix a morpheme added at the end of a word to form a derivative, e.g., -ation, -fy, -ing, -itis.. viii, 13, 14, 20, 22

supervised data data which are assigned labels by human annotators. v, 1

SVD Singular Value Decomposition. 5

term-document matrix a mathematical matrix that describes the frequency of terms that occur in a collection of documents. This is a matrix where. each row represents one document. each column represents one term. 5

tokenized split up into tokens (words). 14–16

topical of or relating to topic. v, 8, 27–30

train the act by which a machine learning classifier uses labeled data to extract patterns with which it may use to better predict the labels of future sets of data. 1

unsupervised data data which are assigned labels automatically by some other method that does not involve human annotation. 2

vector a list of numbers commonly used as computational representations of words or sentences. ix, 5, 17, 23–27

BIBLIOGRAPHY

D. Biber. Dimensions of register variation: A cross-linguistic comparison. 1995.

B. Bloch. Studies in colloquial Japanese II syntax. *Language*, 22(3):200–248, 1946. ISSN 00978507, 15350665. URL http://www.jstor.org/stable/410208.

J. Brooke and G. Hirst. Supervised ranking of co-occurrence profiles for acquisition of continuous lexical attributes. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2172–2183, Dublin, Ireland, Aug. 2014. Dublin City University and Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C14-1205.

J. Brooke, T. Wang, and G. Hirst. Automatic acquisition of lexical formality. In *Coling 2010: Posters*, pages 90–98, Beijing, China, Aug. 2010. Coling 2010 Organizing Committee. URL https://www.aclweb.org/anthology/C10-2011.

P. Brown and C. Fraser. Speech as a marker of situation. 1979.

R. Brown and A. Gilman. The pronouns of power and solidarity. In T. A. Sebeok, editor, *Style in Language*, pages 253–276. MIT Press, Cambridge, Mass, 1960.

Y. Cheng, Z. Gan, Y. Zhang, O. Elachqar, D. Li, and J. Liu. Contextual text style transfer, 2020.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

M. Faruqui and S. Padó. Towards a model of formal and informal address in english. In *EACL*, 2012.

W. Feely, E. Hasler, and A. de Gispert. Controlling Japanese honorifics in English-to-Japanese neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 45–53, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5203. URL https://www.aclweb.org/anthology/D19-5203.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International*

Conference on Acoustics, Speech, and Signal Processing, volume 1, pages 517–520. IEEE, 1992.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

F. Heylighen and J.-M. Dewaele. Formality of language: definition, measurement and behavioral determinants. 1999.

F. Heylighen and J.-M. Dewaele. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7:293–340, 09 2002. doi: 10.1023/A:1019661126744.

J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. *Machine Translation Summit, 2005*, pages 79–86, 2005.

T. Kudo. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge. jp*, 2006.

S. Lahiri. Squinky! A corpus of sentence-level formality, informativeness, and implicature. *CoRR*, abs/1506.02306, 2015. URL `http://arxiv.org/abs/1506.02306`.

R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

F. Luo, P. Li, J. Zhou, P. Yang, B. Chang, X. Sun, and Z. Sui. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5116–5122. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/711. URL `https://doi.org/10.24963/ijcai.2019/711`.

L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

M. Morishita, J. Suzuki, and M. Nagata. Jparacrawl: A large scale web-based english-japanese parallel corpus, 2019.

G. Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

X. Niu and M. Carpuat. Controlling neural machine translation formality with synthetic supervision, 2019.

X. Niu, S. Rao, and M. Carpuat. Multi-task neural models for translating between styles within and across languages. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1086.

U. Okamoto. *Nihongo hyōgen bunten / Okamoto Uichi hen.* Kokusai Bunka Shinkōkai Tōkyō, 1944.

E. Pavlick and A. Nenkova. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 218–224, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1023. URL https://www.aclweb.org/anthology/N15-1023.

E. Pavlick and J. Tetreault. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74, 2016. doi: 10.1162/tacl_a_00083. URL https://www.aclweb.org/anthology/Q16-1005.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

G. D. Prideaux. *The syntax of Japanese honorifics*, volume 102. Walter de Gruyter GmbH & Co KG, 2017.

R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz. JESC: Japanese-English subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1182.

S. Rao and J. Tetreault. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1012. URL https://www.aclweb.org/anthology/N18-1012.

D. Ringel, G. Lavee, I. Guy, and K. Radinsky. Cross-cultural transfer learning for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3871–3881, Hong Kong, China, Nov.

2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1400. URL https://www.aclweb.org/anthology/D19-1400.

A. Romanov, A. Rumshisky, A. Rogers, and D. Donahue. Adversarial decomposition of text representation. *arXiv preprint arXiv:1808.09042*, 2018.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1): 1929–1958, Jan. 2014. ISSN 1532-4435.

Y. Tanaka. Compilation of a multilingual parallel corpus. *Proceedings of PACLING 2001*, pages 265–268, 2001.

Y. Wang, Y. Wu, L. Mou, Z. Li, and W. Chao. Harnessing pre-trained neural networks with rules for formality style transfer. pages 3564–3569, 01 2019. doi: 10.18653/v1/D19-1365.

S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

M. B. Wolfe, M. E. Schreiner, B. Rehder, D. Laham, P. W. Foltz, W. Kintsch, and T. K. Landauer. Learning from text: Matching readers and texts by latent semantic analysis. *Discourse processes*, 25(2-3):309–336, 1998.

R. Xu, T. Ge, and F. Wei. Formality style transfer with hybrid textual annotations. *CoRR*, abs/1903.06353, 2019. URL http://arxiv.org/abs/1903.06353.