Semi-Formal: Semi-supervised formality estimation with Japanese-English parallel corpora

Abstract

The ability to estimate the formality of a sentence in English is useful for a variety of natural language processing tasks such as machine translation, creating appropriately polite dialogue agents, etc. Unlike other languages, such as Japanese, English does not explicitly mark formality and as such generally requires large amounts of labeled data in order to train reliable classifiers. We propose a method of projecting formality labels onto unlabeled English sentences through the use of Japanese-English parallel corpora in order to create a formality data set consisting of over one million labeled sentences, the largest of its kind. Lastly, we show that formality classifiers trained on our data can be fine-tuned with manually labeled data to achieve new state-of-the-art results on relevant shared task benchmarks.

1 Introduction

While the sentences "What's up?" and "How are you?" are very similar in meaning, they differ greatly in terms of their style. These stylistic elements can contain a wealth of information not explicitly available from the text such as the social distance between a pair of speakers (Bramsen et al., 2011), their shared knowledge (Brown and Fraser, 1979), and the goals each have for the interaction (Hovy, 1987). This information, far from being extraneous, may have a larger impact on how the hearer understands the sentence than the literal meaning itself (Hovy, 1987). In this work we will focus on one such stylistic variation – formality.

Robust formality classification has been shown to be useful in a variety of natural language tasks including dialogue systems (Mairesse, 2008; Battaglino and Bickmore, 2015) and sociolinguistic analysis (Danescu-Niculescu-Mizil et al., 2011; Krishnan and Eisenstein, 2015). Despite this, there isn't a particularly well agreed on definition for the term (Pavlick and Tetreault, 2016).

Inf. Japanese	知って何の意味がある？
Inf. English	what good would that have done?
F. Japanese	神が体現する根本を定めているのです
F. English	it defines the primary attributes of divinity

Table 1: Examples of formal and informal sentences from JESC selected by our Japanese regular expression with accompanying English.

Some have defined formality in terms of situational factors such as social distance and shared knowledge (Lahiri et al., 2011) while others have leaned on lexical features such as word choice and part-of-speech (POS) usage (Heylighen and Dewaele, 2002). More recent work has defined formality with a bottom-up approach, allowing individual speakers of a language to each have their own varying definitions of formality (Pavlick and Tetreault, 2016; Lahiri, 2015).

If we accept this definition it becomes necessary to compile formality judgements from many different annotators. Since compiling such annotator judgements for a sufficiently large number of sentences is prohibitively expensive, a semi-supervised approach to formality classification that will allow us to train robust multi-domain classifiers without the need for large amounts of manually labeled data is necessary.

Our contributions are as follows:

- We introduce a method for projecting formality information from Japanese sentences to their English translations.
- We show that this dataset can perform competitively without any manually labeled data.
- We show that augmenting existing manually annotated English formality data with our projected data results in performance improvements for English formality estimation – better than existing semi-supervised datasets.

2 Related Work

Recent work has centered around formality style transfer (Xu et al., 2019; Niu et al., 2018; Wang et al., 2019; Luo et al., 2019), primarily spurred by the release of the Grammarly Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault, 2018). However, underpinning the evaluation of such models is the assumption that we have reliable methods for classifying sentence-level formality which may not necessarily be the case, especially for domains that are very distant from Yahoo Answers.

Additionally, recent research in Formality-Sensitive Machine Translation (FSMT) (Niu et al., 2017) has shown that the addition of formality as an input parameter not only improves the style of output sentences but also gives increases in BLEU score for language pairs such as English/German (Sennrich et al., 2016) and English/Japanese (Feely et al., 2019). These formality-sensitive methods rely on accurate analysis of source language formality at inference time. One recent method of accomplishing this formality estimation is Online Target Inference (Niu and Carpuat, 2019), which infers formality of a given input based on the extent to which it is transformed by a formality style transfer model. While this method shows promise, it depends heavily on the reliability of monolingual style transfer models and the accurate classification of formality.

Several past research efforts have projected elements present in one language onto another language across parallel corpora (Yarowsky and Ngai, 2001). Most relevant to our work is Faruqui and Padó (2012), which investigated the projection of formality onto English using German-English parallel corpora. Their work focused on direct address, as the indicator for formality in German is only found in the different forms of the pronoun "you". An attempt at overcoming this limitation of German-English formality transfer was made by Ringel et al. (2019) through the exploitation of cross-cultural differences and unaligned bilingual corpora, but the proposed method dealt primarily at the document level and not at the level of individual sentences, preventing their algorithm from being used in tasks such as machine translation, and style transfer. Our work seeks to remedy this limitation at the sentence level through the use of Japanese, as Japanese formality determiners are not limited to any one mode of address and are ubiquitous across all forms of the language.

3 Experimental Design

3.1 Formality in Japanese

There are three distinct registers of Japanese formality:

- **Informal**: used primarily with close friends, family, and those who are younger than you.
- **Polite**: used with acquaintances and those of generally equal social status.
- **Formal**[1]: used in addressing those of significantly higher status as well as those whom you want to make feel honored, such as a customer at a check-out counter.

All three of these registers are characterized by distinct sets of verb conjugations, which are easily identifiable through regular expressions. Conveniently, while subjects and objects can be freely dropped from Japanese sentences without being ungrammatical, this is not the case for verbs. This means that every well-formed sentence in Japanese will contain a verb and therefore contain a morpheme for one of these three registers, allowing Japanese-English methods to classify formality for most sentences.

The regular expression we use to classify this formality is a simple string match that operates on raw, un-tokenized Japanese text. If the sentence matches one of our formality keys it is tagged as formal and is otherwise tagged as informal.[2]

In order to evaluate the efficacy of our regular expression, we evaluated a random sample of 100 formal sentences from the Japanese English Subtitle Corpus (Pryzant et al., 2018). As was expected, while our script yielded no explicitly incorrect classifications, there were four sentences that changed speaker mid-sentence or had misaligned or liberal English translations. These sorts of misalignments are particularly common in Japanese-English parallel data due to the vast differences in sentence composition between the two languages.

3.2 Data

In order to properly test our method against other existing forms of formality data augmentation we leverage three main sources.

[1]The formal register in Japanese is occasionally broken up into the honorific register and humble register. For the purposes of this work, these are treated as the same

[2]Combining formal and polite registers is unfortunately necessary for performing binary classification. We leave it to future work to properly take advantage of all three registers of Japanese formality

3.2.1 Japanese English Subtitle Corpus (JESC)

We run our formality regular expression on JESC (Pryzant et al., 2018) to create our semi-supervised formality dataset. JESC was created by scraping data from several repositories,[3] where amateur fan translators can freely upload their own translations for *anime*, *manga*, and television programs. We opted to use this source because of its size (2.8M sentence pairs) and reliable temporal alignments. After running our regular expression, we were able to identify 520K formal and 2.28M informal sentences. We down-sample to create an equal ratio of formal to informal sentences, leaving us with 1.02M sentences, split into 816K train, 102K dev, and 102K test.

3.2.2 Grammarly Yahoo Answers Formality Corpus (GYAFC)

This dataset, introduced by Rao and Tetreault (2018), consists of about 110K formal-informal sentence pairs and has seen significant use in recent work on formality style transfer. It was created by asking human annotators to give formal rewrites to sentences identified as informal by a classifier. For our comparison, we label all formal rewrites as formal and all informal sentences as informal. This results in 222.5K total labeled sentences, split into 178K train, 22K dev, and 22K test.

3.2.3 Pavlick and Tetreault (2016)

This dataset is the largest publicly available source for human formality judgements, totaling 11.1K sentences across four domains (1821 blog, 2775 news, 1701 email, and 4977 yahoo answers). Each sentence is given a grade on the Likert scale (-3 to 3) by 5 annotators. For our purposes, the highest and lowest scores for each sentence were discarded and the middle three scores were averaged and subsequently thresholded at 0 to get a binary formality label. This results in a combined total across all domains of 5560 formal and informal sentences, split into 8816 train, 1216 dev, and 1088 test.

3.3 Experiments

We demonstrate the viability of our generated data for use in formality estimation applications in two major experiments:

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

3.3.1 PT16 Formality Estimation

In this experiment we re-implement the classifier introduced by Pavlick and Tetreault (2016) to serve as a published baseline for comparison. This classifier takes in features such as: named entities, parts of speech, Word2Vec embeddings (Mikolov et al., 2013), punctuation, capitalization, and average formality score as computed by Pavlick and Nenkova (2015). Since this work focuses on binary classification and not categorical, we train a Logistic Regression classifier on these features instead of a Ridge Regression classifier. We use Stanford CoreNLP (Manning et al., 2014) for linguistic processing to model the published classifier as closely as possible. We train a classifier on the training set of each of our main three datasets and evaluate performance on the held-out test set of human data. We compare these results with three very simple baselines.

3.3.2 Fine-Tuning BERT for Formality Estimation

In this experiment we fine-tune BERT to predict sentence-level formality (Devlin et al., 2019), using the HuggingFace transformers library (Wolf et al., 2019).[4] We use the 'bert-base-uncased' model to ensure proper comparison between both spoken and written forms of communication.

We fine-tune a BERT model on five different sets of training data. The first is the 816K JESC training sentences, the second is the 178K GYAFC training sentences, the third is the 8K manually labeled sentences from PT16. The final two are hybrid sets, JESC and GYAFC are each paired with the manually labeled data to see if an improvement in formality classification performance can be achieved through the addition of this extra training data. For these two hybrid data sets, instead of directly combining the two sets of training data we fine-tune first on the larger data set, then on the smaller human data set, both for 2 epochs with a learning rate of 5e-05.

4 Experimental Results

4.1 PT16 Performance on Human Data

The performance of the Logistic Regression classifier trained using PT16 feature vectors is reported in Table 2. Our results show that despite having less than one-tenth the size of GYAFC and less that one one-hundredth the size of JESC the human train

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

Train Data	F1	Acc	Prec	Rec
Human	**77.49**	**76.93**	**80.60**	74.61
JESC	75.42	66.64	62.03	**96.20**
GYAFC	74.56	67.19	63.47	90.33
All Formal	69.47	53.22	53.22	100.0
All Informal	63.75	46.78	46.78	100.0

Table 2: Performance of Logistic regression classifiers trained with input features proposed by PT16 on held out test data of human data against baselines

Classifier	F1	Acc	Prec	Rec
BERT	**70.76**	**71.05**	**71.77**	**69.78**
Logreg	64.27	64.23	63.07	65.52
Neural Net	64.05	62.22	60.11	68.53

Table 3: Performance of JESC models on JESC held out test set of 102K sentences

Data	F1	Acc	Prec	Rec
JESC + H	**82.60**	**81.07**	80.83	84.46
GYAFC + H	82.09	80.79	**81.46**	82.73
Human	80.89	79.50	80.27	81.52
GYAFC	76.52	69.21	64.39	**94.30**
JESC	75.77	68.84	64.63	91.54

Table 4: Performance of BERT models on held out test set of human data; the first column is the training data used.

Classifier	F1	Acc	Prec	Rec
Neural Net	**88.62**	**88.77**	**87.75**	89.52
BERT	87.77	87.50	85.31	90.37
Logreg	87.66	87.29	83.42	**92.34**

Table 5: Performance of GYAFC models on GYAFC held out test set of 22K sentences

data still gives the best performance by over two percent. This portion of the result, makes it clear, at least for the task of sentence-level formality estimation, that high quality labels and appropriate in-domain training data matter much more to performance than the size of the training set.

Taking this conclusion into account makes it all the more surprising that the logistic regression classifier trained on JESC performs comparably well to the classifier trained on GYAFC. The GYAFC data not only has much more consistent and high-quality labels than JESC as in Table 3 and Table 5, but it also ostensibly shares a domain similarity with about 40% of the manually labeled sentences, being both taken from Yahoo Answers.

While we believe further domain-specific investigation into this result is necessary before drawing any conclusions about the ability of GYAFC trained classifiers to generalize to out-of-domain sentences, it is clear that JESC is, at the very least, able to achieve comparable performance to existing data, generalizing well across a variety of domains not found in the primarily dialogue-oriented corpus.

4.2 BERT Fine-Tuning Results

The performance of fine-tuning BERT on each training set is reported in Table 4 We see in this result that we are able to achieve an improvement in formality estimation performance by augmenting existing manually labeled data sets with our generated data. Augmentation of manually labeled data with both GYAFC and JESC show comparable performance increases, with JESC achieving slightly higher accuracy on average. To the best of our knowledge, this is the current state-of-the-art for formality estimation.

Curiously, this result also shows that the BERT model fine-tuned on GYAFC slightly outperforms the model fine-tuned on JESC. While it may be interesting to speculate on reasons why GYAFC might be better suited for adaptation to a BERT model, we hesitate to make any such conclusions as the margin of difference between the two models is so small.

5 Conclusion

The ability to create robust sentence-level formality classifiers that generalize well across many domains would be useful for a variety of natural language understanding tasks. This paper has provided a method for semi-supervised formality transfer which, when used on sufficiently large and well-aligned Japanese-English parallel corpora to generate sentence-level formality data, allows for the training of classifiers that achieve competitive results on shared formality benchmarks. In addition, we provided an augmentation technique that, when used in conjunction with existing manually labeled data, was able to train classifiers that achieve state-of-the-art results. These findings provide important steps towards the practical application of formality classifiers in natural language processing tasks.

References

Cristina Battaglino and Timothy W. Bickmore. 2015. Increasing the engagement of conversational agents through co-constructed storytelling.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782, Portland, Oregon, USA. Association for Computational Linguistics.

Penelope Brown and Colin Fraser. 1979. Speech as a marker of situation. In *Social markers in speech*, pages 33–62. Cambridge University Press.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: Linguistic style accommodation in social media. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 745–754, New York, NY, USA. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Manaal Faruqui and Sebastian Padó. 2012. Towards a model of formal and informal address in english. In *EACL*.

Weston Feely, Eva Hasler, and Adrià de Gispert. 2019. 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. Association for Computational Linguistics.

Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7:293–340.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689 – 719.

Vinodh Krishnan and Jacob Eisenstein. 2015. "you're mr. lebowski, I'm the dude": Inducing address term formality in signed social networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1616–1626, Denver, Colorado. Association for Computational Linguistics.

Shibamouli Lahiri. 2015. Squinky! A corpus of sentence-level formality, informativeness, and implicature. *CoRR*, abs/1506.02306.

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu. 2011. Informality judgment at sentence level and experiments with formality score. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II*, CICLing'11, pages 446–457, Berlin, Heidelberg. Springer-Verlag.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. 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.

François Mairesse. 2008. Learning to adapt in dialogue systems : data-driven models for personality recognition and generation.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Xing Niu and Marine Carpuat. 2019. Controlling neural machine translation formality with synthetic supervision.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. pages 2814–2819.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. 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. Association for Computational Linguistics.

Ellie Pavlick and Ani Nenkova. 2015. 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. Association for Computational Linguistics.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Sudha Rao and Joel Tetreault. 2018. 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. Association for Computational Linguistics.

Dor Ringel, Gal Lavee, Ido Guy, and Kira Radinsky. 2019. 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. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. pages 3564–3569.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Ruochen Xu, Tao Ge, and Furu Wei. 2019. Formality style transfer with hybrid textual annotations. *CoRR*, abs/1903.06353.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.