Automatic Classification of Japanese Formality

Pin-Chen Wang[1] Edison Marrese-Taylor[1,2] Machel Reid[1,3] Yutaka Matsuo[1]
[1]The University of Tokyo, [2]AIST, [3]Google Research
{wangpinchen, emarrese, matsuo}@weblab.t.u-tokyo.ac.jp
machelreid@google.com

概要

In this paper we propose a novel approach to automatically classify the level of formality in Japanese text, using three categories (formal, polite, and informal). We introduce new datasets that combine manually annotated sentences from existing resources, and a great number of formal sentences scrapped from the Japanese Congress. Based on our data, we propose a Transformer-based classification model for Japanese formality which obtains state-of-the-art results in benchmark datasets, as well as on our newly-introduced corpus, showing the effectiveness of our proposed approach.

1 Introduction

Formality or honorifics refers to the level of politeness and respect indicated by the contents, which is a core part of natural language as it indirectly shows how critical the current situation is, one's social rank and status, and also the relation of the people involved in that conversation or any written tasks. However, while many other major languages, including but not limited to English, French, and Mandarin Chinese emphasize formality through the use of standard grammar, more complicated sentence structures (active, passive, use of clauses, etc.), or more advanced and complex choice of vocabulary and phrases, as a society that respects and values etiquette, the Japanese language has its own formality system in order to create specified honorific expressions. This Japanese formality system or Keigo (敬語) is strict and follows a standard grammar format to convert all sentences from one style to another while maintaining the original meaning, word choice, and sentence structure.

Unlike English or other European languages, Japanese formality requires users to identify the status or the relationship with the interlocutor. Japanese honorifics indicate the level of social rank and the hierarchy of the interlocutor or the intimacy one is with that certain person [1]. The crucial point that makes the Japanese formality system stand out from many other major languages is that it allows the users to change the formality level of a sentence from informal to a superiorly high level by merely adjusting the tense of the verb [2]. Generally, Japanese formality can be divided into four different categories [2],

- **Regular (jyotai, 常体)**: a form that is often used in but not limited to a daily conversation with only people one is familiar with or people who are in the equivalent social status.
- **Polite (teineigo, 丁寧語)**: a form that is generally used throughout the whole Japanese society to create some distance between one another. Although this form does not indicate the amount of respect one holds toward others, it helps deliver messages in a polite way that will not be offensive on any occasion.
- **Respectful (sonnkeigo, 尊敬語)**: a form that shows extensive respect, which is used to maximize the preeminence of the interlocutor.
- **Humble (sonnkeigo, 謙譲語)**: a form that specifies humbleness, which is used by the Japanese speakers to minimize their own value in order to highlight the greatness of the interlocutor.

Although grammatically, all the sentences in Japanese can be converted from one formality class to another, the Japanese formality system follows the additional rule [2], where one can always mix the four forms together in one paragraph. The more respectful form one uses in a sentence or a paragraph, the more courtesy one states toward one's interlocutor. Similarly, the more humble form one uses, the more modest one is in the conversation. However, it is also emphasized that when containing too many formal terms in a sentence, the sentence will become annoying and

considered inappropriate regarding the Japanese social rule [2]. Nevertheless, the limitation of the number of formal terms that can be used in a single sentence is so vague that it is impossible to clearly draw a line. The restriction varies depending on the content, length of the sentence, and also the situation of the conversation.

Even if formality control has become a popular task in machine translation, the language models trained for generating appropriate formal sentences are mainly in English or other European languages [3]. To be more specific, although there are numerous English-Japanese parallel corpora for translation tasks as well as formality-classified datasets for some major languages (especially in English), we find a lack of existing resources related to Japanese formality. Moreover, we notice that as generative systems get better at producing sentences in Japanese, we lack mechanisms to automatically determine how well such models can do in terms of adequately generating sentences according to formality levels.

In light of this issue, in this paper, we focus on evaluating and developing resources to work toward a potential solution. We begin by uncovering several flaws, including broken pieces and mislabeling in existing corpora, and introduce a new dataset for Japanese formality that consists of manually labeled sentences (informal, polite, and formal) with examples gathered from some existing sources (IWSLT2022 Special Task on Formality Control for Spoken Language Translation [4] and KeiCO [5]) and also with data extracted from meeting minutes from different committees of the House of Representatives of Japan and the House of Councilors of Japan [1] from the year 1947 to 2022. Then, based on our data, we proposed a Transformer-based classification model for Japanese formality which outperforms any existing classification model for Japanese formality. The models and the corpus are released in this repository [2].

2 Related Work

Communication by way of natural language often includes indicators for respect, sometimes shown via explicit controls for formality. Consequently, for the task of machine translation where the goal is to accurately translate an input sequence into a target language, Formality-Sensitive Machine Translation Model (FSMT) [6], which explicitly takes an expected level of formality as input was introduced as a form of control to make the formality levels consistent. Despite this, research on this task is still limited.

To the best of our knowledge, one of the existing corpora for Japanese formality (the annotated contrastive reference translation dataset) was released as a shared task on Formality Control for Spoken Language Translation announced by the International Conference on Spoken language Translation (IWSLT) in 2022 [4]. This dataset released 1,000 parallel English-Japanese sentences for training and 600 parallel sentences for testing. Another recent paper introduces a corpus only for Japanese formality, where 40 native Japanese volunteers were asked to regenerate and annotate new Japanese sentences based on 3,000 original sentences [5]. The KeiCO corpus contains 10,007 examples and consists of the four forms of the Japanese formality system.

Moreover, current studies rely on human assessment or simple models to verify the performance of formality control of a machine translation model. For example, the FSMT for English-French translation [6] conducted a human study in which they assigned translation pairs for human annotators. Neural Machine Translation (NMT) models for Japanese-English translation [7] and English-German translation [8] depends on rule-based classifiers where they list grammatical rules for the language and classify the input sentence by matching certain syntaxes. In addition, based on the two datasets mentioned above, a BERT-based classifier is proposed with the KeiCO corpus [5] and an XLM-R-based [9] classifier is created for measuring the performance of Pre-trained Multilingual Language Models [10].

3 Proposed Approach

We divide the Japanese language into three categories based on the four formality terms and their corresponding applied situations: (1) Informal (for regular tense), (2) Polite (for polite tense), and (3) Formal (for respectful and humble tenses) — we show examples in Appendix A.

3.1 Datasets

The size and quality of the datasets are vital requisites to maximize the performance of the machine learning models [11]. Therefore, this study uses datasets from three criteria,

1) https://kokkai.ndl.go.jp/#/
2) https://github.com/gg21-aping/Japanese-Formality-Corpus

(1) existing corpus, (2) existing corpus with reannotation, and (3) newly introduced corpus. For annotation for the datasets in (2) and (3), we asked 20 to 30 native Japanese (born and raised in Japan) with ages in the range of 20 to 30 years old to annotate the examples into three classes, informal, polite, and formal. All annotators are currently undergraduate or graduate students of the University of Tokyo, Japan. Furthermore, all the annotations are then double-checked with the Japanese dictionaries by another 3 native Japanese who are also students of the University of Tokyo.

Therefore, we prepare two training sets and two testing sets for all models based on the following dataset mentioned in this section. *TrainV1* consists of 1,200 examples collected from ReIWSLT2022FC, DAILY, and KoKai with 426 informal sentences, 501 polite sentences, and 273 formal sentences, where *testV1* has 300 informal sentences, 330 polite sentences, and 370 formal sentences also collected from the same sources. *TrainV2* consists of 3,500 examples with 1,200 examples from *TrainV1* and another 2,300 examples from the KeiCO corpus, with 989, 1,138, and 1,373 sentences for informal, polite, and formal, respectively. *testV2* consists of 2,001 examples randomly seleceted from the KeiCO corpus, with 531, 503, and 967 sentences for informal, polite, and formal, respectively.

KeiCO [5] We randomly selected 20% of the examples from the original KeiCO corpus as a test set following the original paper [5] and sample 2,300 examples for training purposes.

ReIWSLT2022FC We reannotate a total of 1,000 examples from the dataset introduced by IWSLT for formality control in 2022 [4] as we realized that the Japanese content is considered not reliable. Some of them are broken sentences, and some of them do not carry an understandable Japanese meaning. Besides, the provided dataset is binary classified, which is not as accurate as this paper expected. Also, we discovered after careful reannotation that 44 out of 1,000 examples are mislabeled. After reannotation, 520 sentences are informal, 464 sentences are polite, and 12 sentences are formal.

DAILY In DAILY, this paper randomly collected Japanese sentences from Japanese news, novels, textbooks, business letters, academic documents, etc. The dataset carries well-balanced three classes with 65, 67, and 68 sentences for the informal, polite, and formal classes, respectively. This dataset provides us with a quick glance at how the formality system works in Japanese and helps make up the classification models in the beginning stage of this research.

KoKai Because of the lack of formal sentences in the early stage of the research, this paper has dedicated itself to finding extremely formal examples. As we noticed that politicians in Japan speak in a superior formal way, this paper collected all the meeting minutes from the Japanese Congress from 1947 to 2022. The sentences and phrases used in the committees are considered formal, or at least polite, with very little informal syntax. There are in total 64,630 sentences with 23,672 paragraphs, excluding 11,805 broken sentences which are mostly the names, dates, or titles of the committees or the list of participants. We also assume that the broken sentences and some of the informal sentences scrapped from the meeting minutes are likely to be interrupted sentences, unfinished sentences, or questions during the interpolation. In order to utilize the KoKai corpus, we randomly selected 1,360 examples of the entire 64,630 examples, and have the annotators give labels to the sentences. As a result, we have 137 informal examples, 760 polite examples, and 463 formal examples.

3.2 Models

This paper constructs a rule-based classifier for Japanese formality as the baseline of the study. Then, we built models based on algorithms of logistic regression, naive Bayes, and support vector machine (SVM). Also, as BERT [12] achieved excellent performance in many tasks, we also use our dataset to finetune the *BERT base Japanese* proposed by Tohoku University [13]. We used AdamW as the optimizer and set the learning rate to 1e5. While we use *nagisa* tokenizer [14] for the machine learning classifiers, the transformer model uses the IPA dictionary and tokenizes sequences by the MeCab morphological parser [15].

4 Experimental Setup

Since the performance of machine learning models highly depends on the contents of the training data [11], we questioned the results of the models, suspecting whether the model truly learns the terms of the Japanese formality, or whether the model merely learns to draw a line between the situation implied in the contents.

Therefore, we trained our model with two training sets and tested them on another two testing sets. Formal

sentences in *TrainV1* mainly come from the Japanese Congress, where contents are highly related to politics, education, military, economics, development, etc. On the other hand, formal sentences from *TrainV2* consist of both political text and content from diverse situations. Then, we calculate F1 scores for each model to compare their performances. Besides, we also suspect that batch sizes and the number of epochs may affect the result of the performance. Therefore, for the transformer model, we would also like to try training it with different parameters. Lastly, in order to know whether the models truly work, we then compared the accuracy scores of our models with the performances of the Japanese formality classification models so far.

5 Results

Here, *Transformer V1* represents the transformer-based model which is trained on the training set *TrainV1*, while *Transformer V2* is the transformer-based model trained on the training set *TrainV2*. This paper applies the setup of batch size = 16 and epochs = 20 for transformer-based models as this setup proposed a better performance compared to the others. Full results of the parameter studies are displayed in Appendix B. Almost all the models result in performance greater than the baseline, while the transformer-based model reaches an F1 score of 0.91 for *testV1* and 0.81 for *testV2*. Results of the performance for all the models have been stated in Appendix C

5.1 Perfomance Comparison

To prove that our model is reliable, this thesis compares our performance results with existing Japanese formality classifiers. Table 1 displays the result of comparing the performances tested on the *TEST* set provided by IWSLT 2022 [4] between our model and the binary classifier tested on the *TRAIN* set used for pre-trained models [10]. To also convert our model from a multi-class model to a binary class model, we consider the polite class in our model to become formal. Furthermore, table 2 demonstrates the performance comparison between our model and the classifier based on the KeiCO corpus [5]. The F1 score of both models is calculated based on their performance on the *testV2*. The table suggests that without feeding any examples from the KeiCO corpus to the model during the training process, *Transformer V1* is unable to handle sentences created in the KeiCO corpus. However, after feeding some examples (apart from the testing set) to the training process, *Transformer V2* outperforms the KeiCO classifier with an F1 score of 0.810 over 0.727.

表 1 Comparison of the Performance of the Models testing on IWSLT test set. (The performance given by the original paper [10] is the performance on the TRAIN set.)

Model	F1-score		
	Informal	Formal	Overall
Rippeth+ (2022)	0.98	0.98	0.98
Ours v1	0.97	0.97	0.97
Ours v2	0.97	0.97	0.97

表 2 Comparison of the Performance of the Models testing on KeiCO test set. The "formal" column refers to the accuracy of the model to detect the formal term while the "level" column indicates the performance of detecting the level of honorifics.

Model	F1-score	
	Formality	Hon. Level
Liu+Kobayashi (2022)	0.802	0.727
Ours v1	0.550	0.640
Ours v2	**0.840**	**0.810**

6 Conclusions

Briefly, this paper introduces 2 new datasets (DAILY and KoKai) and a reannotated dataset (ReIWSLT2022) for Japanese formality. Also, we discovered some flaws in the existing dataset that results in unreliable performance. Based on our data, we then report results on mainly 5 models consisting of the knowledge base (rules), algorithms, and neural networks, namely, a rule-based model, logistic regression-based models, naive Bayes-based models, SVM-based models, and transformer-based models. These models carry a state-of-the-art performance that outperforms all the other existing classifiers.

In short, the lack of ability to recognize formal terms, the inaccurate choice of vocabulary and phrases to generate, and the doubtful usage of accurate sentence structure still make it perplexing to achieve the goal of translating appropriate natural languages. Therefore, in light of specifying obstacles in the Japanese formality system, this paper aims to provide detailed and critical instructions to guide further researchers in considering honorifics in Asian societies where etiquette is highly valued.

参考文献

[1] Atsushi Fukada and Noriko Asato. Universal politeness theory: application to the use of japanese honorifics. **Journal of Pragmatics**, Vol. 36, No. 11, pp. 1991–2002, 2004.

[2] 文化審議会. Instruction to japanese formality (敬語の指針). Technical report, Commissioner for Cultural Affairs, Japan, February 2007.

[3] Xing Niu and Marine Carpuat. Controlling neural machine translation formality with synthetic supervision. **Aaai Conference On Artificial Intelligence**, 2019.

[4] Maria Nădejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. CoCoA-MT: A dataset and benchmark for Contrastive Controlled MT with application to formality. In **Findings of the Association for Computational Linguistics: NAACL 2022**, Seattle, USA, July 2022. Association for Computational Linguistics.

[5] Muxuan Liu and Ichiro Kobayashi. Construction and validation of a Japanese honorific corpus based on systemic functional linguistics. In **Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference**, pp. 19–26, Marseille, France, June 2022. European Language Resources Association.

[6] Xing Niu, Marianna Martindale, and Marine Carpuat. A study of style in machine translation: Controlling the formality of machine translation output. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 2814–2819, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[7] Weston Feely, Eva Hasler, and Adrià de Gispert. Controlling Japanese honorifics in English-to-Japanese neural machine translation. In **Proceedings of the 6th Workshop on Asian Translation**, pp. 45–53, Hong Kong, China, November 2019. Association for Computational Linguistics.

[8] Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 35–40, San Diego, California, June 2016. Association for Computational Linguistics.

[9] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco GuzmÃ¡n, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. **arXiv preprint arXiv: Arxiv-1911.02116**, 2019.

[10] Elijah Rippeth, Sweta Agrawal, and Marine Carpuat. Controlling translation formality using pre-trained multilingual language models. In **Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)**, pp. 327–340, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics.

[11] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. **Foundations of Machine Learning**. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2018.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[13] Tohoku University,. Bert base japanese. https://huggingface.co/cl-tohoku/bert-base-japanese, 2019.

[14] 池田大志. nagisa: Rnn による日本語単語分割・品詞タグ付けツール. https://qiita.com/taishi-i/items/5b9275a606b392f7f58e, 2018.

[15] Takumitsu Kudo. Mecab : Yet another part-of-speech and morphological analyzer. 2005.

A Classification Classes

For example, given an original English sentence "*I would like to inquire about the schedule of the final exam. We had this conversation earlier that the schedule for my thesis defense overlapped with the exam. I would like to know if there is a possibility that I can take the exam one week after the official exam date.*", "期末試験の日程を聞きたい。修論発表の日程が試験と被っている話を前したのだが、正式な試験日の一週間後に受ける形で対応してもらいたい。よろしく。" is considered informal, "期末試験の日程について聞きたいです。修論発表の日程が試験と被っている件について以前話したのですが、正式な試験日の一週間後に受験するという形で対応してもらえますか。よろしくお願いします。" is considered polite, and "期末試験の日程について伺いたいことがございます。修論発表の日程が試験と被っている件について以前お話しさせていただいたのですが、正式な試験日の一週間後に受験させていただくという形で対応していただくことは可能でしょうか。ご検討よろしくお願いいたします。" is considered formal. Colored phrases refer to the tense of the verb for the different categories.

B Hyper-Parameter Tuning

表 3 Performance on Transformer V1 with different batches and epochs.

Transformer v1	F1 scores			
	informal	polite	formal	avg.
batch = 16 epoch = 20	**0.95**	**0.89**	**0.87**	**0.90**
batch = 32 epoch = 20	0.94	0.87	0.85	0.89
batch = 64 epoch = 20	0.94	0.87	0.84	0.88
batch = 128 epoch = 30	0.94	0.88	0.85	0.89
batch = 256 epoch = 40	**0.95**	0.88	0.86	**0.90**

In the case of finding out if adjusting parameters affect the performance result on our dataset when finetuning the transformer-based model, different parameters were fed when training *Transformer V1*. Table 3 shows the results on the performance of models with different batch sizes and epochs. The size of the training set might be a core factor that results in such F1 scores for each setup. However, a simple hypothesis can be made that compared to polite and formal sentences, it is considerably easier for a model to recognize the informal tense of the Japanese language compared to other honorific levels.

表 4 Performance on all models trained by different versions of the training set.

Model	Train	F1 scores (test v1 / test v2)			
		informal	polite	formal	avg.
Rule	-	0.95 / 0.87	0.88 / 0.56	0.86 / 0.48	0.90 / 0.62
LR	v1	0.90 / 0.77	0.81 / 0.53	0.77 / 0.55	0.83 / 0.61
LR	v2	0.92 / 0.85	0.82 / 0.87	0.80 / 0.64	0.85 / **0.80**
NB	v1	0.82 / 0.52	0.71 / 0.44	0.58 / 0.22	0.71 / 0.39
NB	v2	0.87 / 0.74	0.65 / 0.60	0.71 / 0.25	0.74 / **0.63**
SVM	v1	0.91 / 0.79	0.82 / 0.54	0.77 / 0.54	0.84 / 0.62
SVM	v2	0.94 / 0.88	0.83 / 0.67	0.79 / 0.84	0.86 / **0.81**
Ours v1	v1	0.95 / 0.83	0.89 / 0.56	0.87 / 0.55	0.90 / 0.64
Ours v2	v2	0.95 / 0.87	0.89 / 0.67	0.88 / 0.84	**0.91** / **0.81**

C Different Models

Table 4 displays the accuracy achieved by each model with different training sets and being tested with different testing sets. The results demonstrate that overall, all models perform better than the baseline model. However, the average F1 scores on different testing set significantly diverge, for *testV1*, the F1 scores moderately increase when *TrainV2* is used. When it comes to *testV2*, the accuracy scores remarkably rocket when training on *TrainV2*. Recall that almost all the formal sentences in *TrainV1* come from the Japanese Congress, formal examples are highly politics-related. Hence, this thesis makes the assumption that *Transformer V1* learns how to determine whether a given input is of a formal content rather than with a formal Japanese term.