Towards better evaluation for Formality-Controlled English-Japanese Machine Translation

Edison Marrese-Taylor[1,2], Pin-Chen Wang[2], Yutaka Matsuo[2]
[1] National Institute of Advanced Industrial Science and Technology
[2] Graduate School of Engineering, The University of Tokyo
{emarrese,wangpinchen,matsuo}@weblab.t.u-tokyo.ac.jp

Abstract

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 a new dataset that combine manually-annotated sentences from existing resources, and formal sentences scrapped from the website of the House of Representatives and the House of Councilors of Japan. Based on our data, we propose a Transformer-based classification model for Japanese, which obtains state-of-the-art results in benchmark datasets. We further propose to utilize our classifier to study the effectiveness of prompting techniques for controlling the formality level of machine translation (MT) using Large Language Models (LLM). Our experimental setting includes a large selection of such models and is based on an En→Ja parallel corpus specifically designed to test formality control in MT. Our results validate the robustness and effectiveness of our proposed approach and while also providing empirical evidence suggesting that prompting LLMs is a viable approach to control the formality level of En→Ja MT using LLMs.

1 Introduction

Communication by way of natural language often includes indicators for respect to acknowledge the hierarchy, interpersonal relationship, and power dynamics of the participants in a conversation or written text. In this context, formality or honorifics refers to the set of linguistic features used to establish the degree of respect and deference conveyed in a given context.

Naturally, these phenomena exhibit significant variation across different languages and cultures (Biber and Conrad, 2019). While many European languages 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, the Japanese language has its own formality system. This system, named Keigo (敬語), requires users to identify the status or the relationship with the interlocutor, is strict, following a standard grammar format (Fukada and Asato, 2004), and can generally be divided into four different categories (Aoki et al., 2007), as follows.

- **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 (sonkeigo,** 尊敬語**)**: a form that shows extensive respect, which is used to maximize the preeminence of the interlocutor.
- **Humble (kenjyougo,** 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.

In this context, what makes Japanese formality stand out is that it allows to convert any sentences from one style to another by simply adjusting the tense of the verb (Aoki et al., 2007), while maintaining the original meaning, word choice, and sentence structure.

Additionally, the system follows one additional rule (Aoki et al., 2007), 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

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 551–560
December 6–7, 2023.

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 in Japanese social rules (Aoki et al., 2007).

Given the importance of formality in language generation systems such as machine translation (MT), the ability to control formality and honorifics is a critical factor in achieving accurate and appropriate results. In particular, for the Japanese language, failure in recognizing and incorporating levels of formality can result in unnatural, impolite, or disrespectful translations, which can impede effective communication across diverse linguistic and cultural contexts (Fukada and Asato, 2004). We therefore think that developing and refining MT models that can accurately control honorific levels is crucial for this language. Although formality-controlled machine translation (FCMT) has gained popularity for languages like English (Niu and Carpuat, 2020), there is a substantial lack of resources to tackle the formality problem for Japanese, which extends to the more fundamental task of formality detection.

In light of this issue, we focus on developing resources to improve formality detection in Japanese. We begin by uncovering several flaws on existing corpora for the task, including issues such as the presence of ungrammatical sentences, as well as wrong formality labels. To alleviate these issues, we introduce new resources for Japanese formality detection which consists of manually-labeled sentences annotated with three formality classes (informal, polite, and formal). We propose this three-way setting in opposition to existing resources which are annotated using binary labels, to better approximate the nature of formality of the Japanese language. As existing resources (Nadejde et al., 2022; Liu and Kobayashi, 2022) lacked data for the formal label, a part of our dataset is constructed with sentences sampled from these sources and with text obtained from meeting minutes from committees of the House of Representatives and the House of Councilors of Japan [1].

Furthermore, as language generation models based on Large Language Models (LLMs) have recently been able to attain substantial performance improvements on language generation benchmarks, we note that the lack of a consistent evaluation method makes it difficult to verify to what extent such models can perform formality control. In MT, current studies mainly rely on human assessment or simple models. For example, Feely et al. (2019) and Nadejde et al. (2022) use rule-based methods where lists of grammatical rules are combined with pattern-matching to perform classification. Though formality-level classifiers for Japanese based on machine learning have been proposed in the past (Rippeth et al., 2022; Liu and Kobayashi, 2022), so far this has been without focus on MT or lacked proper evaluation.

In consideration of the above issue, in this paper we propose a novel approach, based on machine learning, to evaluate the ability of En→Ja MT models to perform formality-control. Concretely, we use our dataset to train a robust Transformer-based classifier that leverages a masked-language model, which is able to obtain state-of-the-art performance on our dataset and on existing Japanese formality detection benchmarks. Following recent work relying on machine learning models to evaluate language generation, such as BERTScore (Zhang et al., 2019), and MT models, such as COMET (Rei et al., 2020a), we present an empirical study using our classifier to evaluate the zero-shot ability of several state-of-the-art LLMs to perform formality control.

Our results validate the effectiveness of our proposed approach and show that, compared to existing evaluation techniques that rely on rules and expression-matching, it offers a robust, reliable, and accurate evaluation metric for formality-controlled MT systems. We further demonstrate the ability of LLMs to generate sequences with varying levels of formality through the use of well-designed prompts, concretely showing that both GPT-3 and ChatGPT can attain a formality control accuracy of approximately 90%, and ultimately suggesting that prompting LLMs can result in better formality control performance than fine-tuned MT models. We release or data and trained models[2] to encourage further research on this topic.

2 Related Work

To the best of our knowledge, previous work on formality detection for Japanese is relatively recent and limited in scope, with only two existing resources. On the one hand, we find the Japanese portion of the CoCoA-MT (Nadejde et al., 2022)

[1] https://kokkai.ndl.go.jp/

[2] https://github.com/epochx/japanese-formality

dataset, which was released for the 2022 Shared Task on Formality Control at IWSLT (Anastasopoulos et al., 2022) and contains a total of 1,600 parallel English-Japanese sentences (1,000 for training, and 600 for testing). The source data for this corpus comes from Topical-Chat4 (Gopalakrishnan et al., 2019), as well as Telephony and Call Center data, containing text-based conversations about various topics. For each segment, one reference translation for each formality level (formal and informal) were collected. For the Japanese translations, informal was mapped to jyoutai, and formal was mapped to teineigo, sonkeigo and/or kenjyougo.

On the other hand, we find the recently-released KeiCO corpus (Liu and Kobayashi, 2022), which has a total of 10,007 examples across the four forms of the Japanese formality system (Levels 1 to 4, according to the paper). It additionally contains detailed information about the presence of level-related honorifics —a sentence may contain markers for multiple levels of politeness— the social relationship between the speaker and the listener, and conversational situations or topics. To obtain this data, 40 native Japanese volunteers were asked to regenerate a total of 3,000 sentences coming from machine translation, dialogue systems, and semantic analysis systems, by filling in blanks with honorifics.

The two datasets mentioned above have been used to train Transformer-based classifiers. Liu and Kobayashi (2022) rely on Japanese-BERT (Suzuki and Takahashi, 2021), while the submission of Rippeth et al. (2022) for the 2022 Shared Task on Formality Control at IWSLT relied on XLM-R (Conneau et al., 2020).

Our work is also related to FCMT. In this context, recent approaches have relied on formality-annotated parallel corpora such as CoCoA-MT, early work on this task resorted to other resources such as rule-based generation of synthetic data for English-Japanese (Feely et al., 2019) and English-German (Sennrich et al., 2016), as well as synthetic supervision by means of multi-tasking (formality classification and machine translation). We also find that these studies rely on rule-based simple approaches to measure the accuracy of formality control in the translation, or directly perform human assessment. For example, the FSMT approach English-French by Niu et al. (2017) conducted a human study in which they assigned translation pairs for human annotators. Neural CFMT models for English-Japanese (Feely et al., 2019) and English-German (Sennrich et al., 2016) depend on rule-based classifiers, where grammatical rules for the language are listed and matched.

The recent rise of LLMs has enabled models to perform certain language generation tasks in zero-shot or few-shot manner (Brown et al., 2020), or by means of prompts. Some of these capabilities have been further enhanced by means of prompt-based training (Sanh et al., 2022), where zero-shot generalization is induced by explicit multitask learning. This work is relevant to our paper, as we test the ability of several such models to perform zero-shot FCMT. Our study also considers multilingual efforts in Neural MT, admittedly also a kind of LLM, where we look at M2M100 (Fan et al., 2021) and NLLB200 (Costa-jussà et al., 2022)

Finally, we also find recent work on using few-shot prompting-based techniques to control the formality level of English-German Machine Translation (Garcia et al., 2023). Also, Pu and Demberg (2023) recently performed an in-depth study of the capabilities of ChatGPT to generate text in different styles, including formal/informal labels, showing that the model sometimes incorporates factual errors or hallucinations when adapting the text to suit a specific style.

3 A robust classifier for Japanese Formality

3.1 Data

The size and quality of the datasets are vital requisites to maximize the performance of the machine learning models (Mohri et al., 2018). As one goal of our work is to train a robust classifier for Japanese formality, we look at two main issues. In contrast to existing resources, which either offer limited flexibility by simplifying the dynamics of Japanese formality into two classes (Nadejde et al., 2022), or are too specific by exactly following the grammar (Liu and Kobayashi, 2022), we propose a compromise between these and divide the Japanese language into three categories based on the four formality levels and their corresponding applied situations: (1) “Informal” (for regular tense), (2) “Polite” (for polite tense or teineigo), and (3) Formal (for respectful and humble tenses). Below, we detail how we transformed existing datasets for our purposes, created new resources when necessary, and how we constructed a final curated corpus to train our model.

RECOCOA-MT As we divided Japanese formality into 3 classes, this suggested that the re-utilization of the Japanese portion CoCoA-MT corpus required a transformation of the labels, so we began by analyzing the data. During this stage, we found that many of the examples of the parallel corpus contain broken sentences, while in many cases other sentences do not have an understandable Japanese meaning. Based on these observations, we decided to re-annotate the data and recruited volunteer Japanese native speakers to proceed[3]. During the re-annotation procedure, we confirmed that 44 out of the 1,000 training examples were mislabeled. After re-annotation and filtering, 520 sentences are labeled as informal, 464 sentences as polite, and 12 sentences as formal.

KOKAI As seen above, the re-annotation of COCOA-MT showed that the label distribution in this dataset was heavily skewed away from the formal label, which suggested that more data for this particular level of formality was required. Noting that Japanese political committees tend to rely on language that is considered formal, or at least polite, with very little informal syntax, we proceeded to collect all the meeting minutes from the Japanese Congress (House of Representatives of Japan and the House of Councilors of Japan) from 1947 to 2022. In total, we obtained 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 surmise some of these broken sentences, as well as the informal sentences that we observed upon close examination, are likely interrupted utterances that occurred during the sessions. Despite the overall formal nature of the source of data, to ensure the quality of the labels we use for training, we randomly selected 1,360 examples from the raw data and ask our volunteer Japanese native speakersfootnote:annotators to annotate the examples following the same procedure as before. As a result, we obtain 137 informal examples, 760 polite examples, and 463 formal examples.

DAILY We collected 200 sentences sampled from Japanese news, novels, textbooks, business letters and academic documents. The dataset is well-balanced across our three labels with 65, 67, and 68 sentences for the informal, polite, and formal classes, respectively. We use this small corpus mainly for preliminary experiments, but also include these examples in the data used to train our model, as explained below.

KEICO We note that according to Liu and Kobayashi (2022), both respectful and humble tenses are used for the proposed formality Levels 1 and 2. We therefore simply map these two classes to our formal class, to make the annotations compatible with our setting

Using these sources of data except the KEICO, we prepared a first training set consisting of 1,000 examples (426 informal, 501 polite sentences, and 273 formal), leaving a total of additional 200 examples left for development purposes. Though KOKAI has been collected to specifically cover for the lack of annotated data for the formal label in RECOCOA-MT, because the contents are highly related to politics and other related domains, we hypothesize that by only utilizing examples derived for this dataset for training may lead to models that may fail to generalize well to other situation where the respectful and humble tenses are also utilized. To that end, for the initial training set we purposely omit examples from KEICO, which contains formal examples from a diverse set of domains, and build a second training set that relies on examples taken from this corpus to balance the topic distribution. We take a total of 2K examples from KEICO (with 530, 503, and 967 sentences for informal, polite, and formal classes, respectively.)

3.2 Model

Drawing from the success of classifiers based on BERT (Devlin et al., 2019), which have achieved excellent performance in a large selection of downstream tasks, and following Liu and Kobayashi (2022), we propose to use Japanese-BERT (Suzuki and Takahashi, 2021) to train a formality classifier for Japanese formality. The input text is preprocessed and tokenized using the MeCab morphological parser (Kudo, 2005), which is what Japanese-BERT utilizes. For training, we used the AdamW (Loshchilov and Hutter, 2019) optimizer, with a learning rate of 10^{-5}, a batch size of 16 and, and train for a maximum of 20 epochs.

We evaluate our model in the test portions of existing datasets, namely, COCOA-MT and KEICO.

[3]We recruited 30 native Japanese speakers within 20 and 30 years old. All annotators are currently undergraduate or graduate students of a university in Tokyo, Japan. Furthermore, these annotations were double-checked with the help of Japanese dictionaries by 3 additional native Japanese speakers who are graduate students of the same university.

For the former, our analysis reveals that out of 600 examples in the test set, only 594 have been made available, which we utilize in our study. For the latter, since no official test splits are provided, we try to follow the experimental setting proposed by Liu and Kobayashi (2022)[4] and randomly selected 20% of the examples to test.

To contextualize our contributions and put the performance of our classifier in context, we consider a selection of baselines taken from previous work, as well as our implementations of newly-introduced models, and use F1-Score for evaluation.

On CoCoA-MT, we consider the rule-based classifiers proposed by Nadejde et al. (2022) and Feely et al. (2019), as well as our Transformer-based classifier. To test our model in this dataset, which is a binary classification setting, we either train another Transformer-based model on binary labels, or simply convert the prediction of the 3-way models into binary classification by considering all the other 3 tenses except for regular form as formal.

For this dataset, we additionally propose a new rule-based classifier for Japanese formality, which we adapt from Feely et al. (2019). Concretely, we propose ways to mitigate some limitations that were identified in the existing model. For example, the original rules assigned "ない (negative - present tense)" and "なかった (negative - past tense)" to the polite class, while we consider that both of them should be the regular form. In our approach, we label all sentences that are not classified as belonging to the "Polite" and the "Formal" class as "Informal".

We omit results by Rippeth et al. (2022), who fine-tune XLM-R on binary classification between formal and informal classes, but only report accuracy on the development set, defined as the last 50 paired contrastive examples from each language, which we regard as too small and incompatible with our setting. Their model obtains an accuracy of 98% on both the formal and informal classes on this set.

For the KEICO dataset, we compare the performance of our Transformer-based and rule-based classifiers against the BERT-based classifier proposed by Liu and Kobayashi (2022). Since this classifier is trained on a different label set compared to our approach, we proceed as follows: (1) we compare the average F1-score for the respectful and humble term detection task in (Liu and Kobayashi, 2022) against the F1-score of our classifier on the formal label, which we regard as a roughly-equivalent setting, (2) as our approach directly collapses Levels 1 and 2 in Liu and Kobayashi (2022) to our formal label, while Level 3 (which uses teineigo) and Level 4 (no honorifics) perfectly match our polite and informal class, respectively, we compare F1-scores as-is against the overall classification performance.

[4]Their reported metrics are the result of 10 runs with different initialization, and each time 20% of the examples are randomly chosen for the evaluation.

Model	Precision	Recall	F1-score
Nadejde et al. (2022)	0.70	0.49	-
Feely et al. (2019)*	0.87	0.83	0.83
Rule-based*	**0.98**	**0.98**	**0.98**
no KEICO samples			
Transformer 3-way*	0.97	0.97	0.97
Transformer 2-way	0.97	0.97	0.97
with KEICO samples			
Transformer 3-way*	0.97	0.97	0.97
Transformer 2-way	0.97	0.97	0.97

Table 1: Performance formality-level classifiers for Japanese on CoCoA-MT, where * indicates models that were originally designed for 3-way classification, but adapted for binary formality labels by considering polite, respectful, and humble forms as Formal. Precision and recall values from (Nadejde et al., 2022) are based on M-Acc score, and are computed on a 300-example subset of the data. F1-scores were not reported, so we omit them.

3.3 Results

As can be seen in Table 1, both our rule-based and Transformer-based models are able to outperform previous work on CoCoA-MT by substantial margins. We further notice that both models are able to attain very similar, and extremely high performance of 97% F1-score, and that neither the change in label setting, nor the addition of examples from KEICO have any effect on the performance. We think these results are compelling evidence suggesting the limited quality of the examples in this dataset. Based on this, we recommend researchers to consider other benchmarks instead.

Table 2 shows our results on the KEICO dataset. We see that our Transformer-based classifier obtains an overall F1-score of 0.84, surpassing of the classifier proposed by Liu and Kobayashi (2022). By contrast, our rule-based classifier only obtains

Model	F1-score	
	Formality	Hon. Level
Liu and Kobayashi (2022)	0.802	0.727
Rule-based	0.620	-
no KEICO samples Transformer	0.550	0.604
with KEICO samples Transformer	**0.840**	**0.810**

Table 2: Summary of results on the KeiCO dataset. The "Formal" column refers to the accuracy of the model to detect formal terms, while the "Level" column indicates the performance of detecting the level of honorifics.

an F1-score of 0.620, showing that rule-based methods, as comprehensive as they may be, offer limited reliability in multi-domain scenarios.

We also observe that the addition of examples from the KEICO dataset to the training data has a substantial impact on the performance of our model. It is only when these examples are added that our Transformer-based model is able to outperform the baseline. We think this result validates our domain-shift hypothesis, suggesting that examples from KOKAI offer only a narrow variety of expressions of Japanese formality, which do not allow the model to generalize well to more general domains.

Overall, our results suggest that the KEICO dataset offers a more compelling and real-like arena to evaluate the accuracy of Japanese formality classifiers.

4 Empirical Study

Having demonstrated the abilities of our Transformer-based classifier of Japanese formality, we now turn to a more practical issue, and tasks ourselves with testing the proposed approach in a real scenario. We examine formality abilities of English to Japanese machine translation using a zero-shot prompting approach. To the best of our knowledge, our work is the first one to study this issue.

4.1 Experimental Setup

Data We utilize the COCOA-MT En→Ja test set for our experiments. As mentioned earlier, examples in this dataset exhibit numerous flaws, including incomplete and semantically meaningless sentences, but since no other suitable dataset exists, we resort to this dataset nonetheless. We assume the existence of tuples $(x, y_{\text{formal}}, y_{\text{informal}})$ where x is the input sentence in English, and y are the target sentences in Japanese at different formality levels. Using the original 594 English sentences, below we show how we prompt our selection of models to produce both informal and formal Japanese translations.

Models We utilize large multilingual MT models trained on massive parallel corpora, specifically, M2M100 (Fan et al., 2021) and NLLB200 (Costa-jussà et al., 2022). Additionally, we experiment with M2M100 models of different sizes, including the 418M and 1.2B models. For each MT model, we use the English sentences from COCOA-MT as input, and concatenate them with a prefix prompt $p \in P = \{\text{formal}, \text{informal}\}$ which is added using square brackets. Thus, the input to the models is expressed as "$[p]; x$," where ; denotes white-space-based concatenation.

Moreover, as LLMs have shown good performance on MT when provided with an appropriate prompt, we also experiment using GPT-3 (Brown et al., 2020) and ChatGPT. We use similar prompts to those used for the MT models, but suggest more clearly to the models to perform the formality control task by using "Translate English to p Japanese: x". For ChatGPT, as the official API was not yet available at the time of our experiments, so we manually input a total of 1,188 examples (594 examples for each informal and formal setup) into the web client of ChatGPT Plus[5]. We also consider the recently-released llama2 models (Touvron et al., 2023), specifically the chat versions, which have been optimized for dialogue. We utilize the 7B-parameter and 13-B models, the latter we quantize to 4-bits using QLoRA (Dettmers et al., 2023) in order to fit our GPU memory. We follow the approach by the original paper to create our prompt, and test two settings (1) a zero-shot approach where the model is directly asked to generate translation, and (2) a one-shot setting, where we incorporate a source-target translation example for the given formality target. We construct this example manually, making sure it has minimum overlap with the examples from our data.

Finally, we also consider the Transformer-based model by Nadejde et al. (2022) as a baseline. This model is a 20-layer encoder and 2-layer decoder Transformer trained from scratch on the COCOA-MT, with the help of data augmentation techniques.

[5] https://openai.com/blog/chatgpt

Model	Cmp (%)	COMET	BLEU	M-Acc	Accuracy		
					Rule	T-3	T-2
Nadejde et al. (2022)	100	-	22.20	0.76 (-)	-	-	-
M2M100 (418M)	100	0.731	16.19	0.49 (0.18)	0.51	0.49	0.51
M2M100 (1.2B)	100	0.744	17.25	0.50 (0.19)	0.50	0.49	0.49
NLLB200 (600M)	100	0.733	8.83	0.47 (0.19)	0.49	0.48	0.48
llama2-chat (7B)	74.8	0.698	8.53	0.52 (0.24)	0.55	0.54	0.54
+ one shot	84.1	0.617	6.76	0.51 (0.26)	0.56	0.51	0.56
llama2-chat (13B)	55.4	0.731	11.36	0.83 (0.19)	0.64	0.63	0.63
+ one-shot	91.3	0.561	8.05	0.61 (0.30)	0.56	0.56	0.57
GPT-3	100	0.875	**23.79**	**0.86 (0.25)**	**0.91**	0.90	0.90
ChatGPT	98.9	**0.868**	20.63	0.83 (0.25)	**0.91**	**0.91**	**0.91**

Table 3: Performance of our experiments with formality-controlled En→Ja MT, including results MT models (Nadejde et al., 2022) fine-tuned on the data, and zero-shot approaches using pre-trained MT models and LLMs. Here, T-3 and T-2 indicate the proposed 3-way and binary Transformer-based classifiers, and Cmp. is short for Compliance, showing the percentage of output that contained valid translations. For M-Acc (Nadejde et al., 2022), we also show the coverage of the matched sentences between parenthesis, as this evaluation metric model overlooks examples that do not match its rules.

Evaluation We perform evaluation in terms of the quality of the generated translations, and in terms of the ability to perform formality control. For the former, we follow previous work and report and BLEU scores (Papineni et al., 2002) relying on the sacrebleu[6] implementation (Post, 2018), and also consider COMET (Rei et al., 2020b) using the "wmt22-comet-da" model, which has multilingual support. For the latter, we rely on Matched-Accuracy (M-Acc) (Nadejde et al., 2022) which is a rule-based corpus-level metric for CoCoA-MT. M-acc works by checking if the hypothesis contains: a) any of the formality-marking phrases annotated in the formal reference and b) none of the phrases annotated in the informal reference (or vice versa). Crucially, sentences that are not matched are simply skipped. This metric was shown by Nadejde et al. (2022) to be relatively reliable for Japanese, obtaining a precision and recall of 0.7 and 0.49, respectively, when tested on a random sample of 300 sentences that were manually annotated by two professional translators. Finally, we utilize our proposed rule-based and Transformer-based classifiers. Finally, we also measure the zero-shot or few-shot ability of LLMs to "comply" with the given prompt by generating plausible translations. Based on the provided instruction, we use heuristics to parse and extract the translation from the text generated, and report the percentage of output that our heuristics are able to parse successfully.

4.2 Results

Table 3 summarizes our results on the formality control in En→Ja MT performance of all the models considered. We see that zero-shot prompting techniques work much better on LLMs than on pre-trained multilingual MT models, with the former attaining the best performance overall. In particular, we see that zero-shot techniques based on prompting lead to substantially low BLEU scores and formality control accuracy when tested on pre-trained multilingual MT models, which are also outperformed by the fine-tuned models by (Nadejde et al., 2022). This suggests that pre-trained multilingual MT models may simply lack the ability to be prompted for formality control.

In terms of model compliance, we notice that prompting LLMs leads to unstable behavior, with models often not following the provided instruction. This therefore leads them to not generate a valid translation, or to do so in a what such that it is not feasible to find the translation automatically in the model output. For example, some models do not follow the input-output pattern described in the prompt, while others tend to explain their translations in some cases. Finally, llama2 models sometimes refused to provide a translation for

[6] https://github.com/mjpost/sacrebleu

safety reasons, as they detected words regarded as rude or potentially harmful in the input.

Moreover, our results shed light on the reliability issues of M-Acc which, due to its hard matching approach, ends up ignoring many of the translations generated by the systems we test. We observe that across all our tested systems, its coverage lies between 0.18 to 0.25. M-acc is in principle designed to work only for the CoCoA-MT dataset. While this is allegedly a strong limitation already, we think the coverage issue we observed suggests that the approach may be even more limited.

In contrast to these results, we observe that both our Transformer-based and rule-based approaches offer no coverage issues, while also agreeing with each other and with the overall M-acc scores. We think these results validate our techniques as valid alternatives for the evaluation of formality-controlled MT, setting a potential direction for future developments.

5 Conclusions

This paper explores new alternatives to evaluate the ability of En→Ja MT models to perform formality control, proposing classifiers based on rule-based methods and a machine learning approach using HuggingFace Transformers[7] (Wolf et al., 2020).

To build robust models, we focus on developing resources to improve formality detection in Japanese, uncovering several flaws on existing corpora for the task, and introducing new annotated datasets. In contrast to prior work approaching formality using binary labels, we use three classes (informal, polite, and formal) to better approximate the ways honorifics are used in the Japanese language. Extensive experiments on benchmark datasets show that our proposed models offer state-of-the-art performance.

Finally, we empirically show that our machine-learning approach is superior to existing evaluation techniques for formality-controlled MT systems, offering a reliable and accurate evaluation solution. The study also demonstrates the ability of LLMs to generate sequences with varying levels of formality through well-designed prompts, resulting in state-of-the-art results in En→Ja formality-controlled MT. Our findings provide a valuable contribution to the NLP field by presenting a new approach to evaluate formality-controlled MT systems and highlighting the effectiveness of LLMs in this task.

[7] https://huggingface.co/docs/transformers

Limitations

In this work, we have introduced both data and models to tackle the task of formality detection in the Japanese language. Though our results suggest that we have been able to build a robust classifier that obtains good performance, we offer no empirical evidence to suggest how well these capabilities could generalize to untested domains or situations.

Moreover, as some of our experiments involved black-box models that are only accessible through an API, such as GPT-3 and ChatGPT, we are unable to offer reliability in replicating those results. Upon acceptance, we will be releasing the output we obtained from these models for the sake of reproducibility of our experiments.

Finally, we also utilize pre-trained models either as baselines or to initialize our proposed classifier, and we think this is an important driver of the performance we observed. This may be an issue where access to pre-trained models is limited.

References

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Tamotsu Aoki, Yoshimitsu Aoyama, Takashi Atoda, Yoshiaki Ishizawa, Danjuro Ichikawa, Emi Uehara, Fumiko Okada, Tadaaki Odaka, Tsuneaki Kawamura, Yasuko Tabata, Takako Tamura, Hideki Tomizawa, Nakayama Nobuhiro, Kazuo Nishi, Suzuko Nishihara, Toyohiro Nomura, Tomihi Maeda, Kazuko Matsuoka, Mayumi Mori, and Nobuo Monya. 2007. Instruction to Japanese Formality (敬語の指針). Technical report, Agency for Cultural Affairs, Government of Japan (文化審議会).

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48.

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.

Atsushi Fukada and Noriko Asato. 2004. Universal politeness theory: application to the use of Japanese honorifics. *Journal of Pragmatics*, 36(11):1991–2002.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10867–10878. PMLR.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Takumitsu Kudo. 2005. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.jp*.

Muxuan Liu and Ichiro Kobayashi. 2022. 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*, pages 19–26, Marseille, France. European Language Resources Association.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8568–8575.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. 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*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Dongqi Pu and Vera Demberg. 2023. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Masatoshi Suzuki and Ryo Takahashi. 2021. Japanese bert. `https://github.com/cl-tohoku/bert-japanese`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.