Multi Politeness-Domain Neural Machine Translation for Japanese and Korean

Henry Li Xinyuan, Jerry Chen, Ray Lee

Autumn 2021

1 Training Data

1.1 Choice of Corpus

Countless corpuses of Japanese exist on the Internet, yet the ones that would be suitable for our needs are far and few between. There is typically a strong correlation between formality and context, which is not bad news for us since relying purely on morphology to label formality would have problems of its own. However, we want to avoid introducing into our corpus large chunks of sentences with the same context in the same formality domain, lest any of our models learns to classify contexts rather than formality. Many such examples of bad corpuses exist, such as the corpus of Japanese legal documents: in Japanese, all legal documents are written in informal form (contrary to what one might assume); we must be very careful when using such corpuses by balancing and mixing them with corpuses from other sources and with different formality domains. Examples of good corpuses include the subtitle corpus, although the translation quality of some of the sentences in that corpus has been questioned. In general, Japanese is a more context-based language, and it often leaves out parts of the phrase than can be inferred. In most cases, this means omitting the subject. This means that English translation will generally have more context than the Japanese equivalent. Currently, there are no corpuses that avoid this issue, as any such corpus would have to be either laboriously picked out or have a heavy bias toward formal sentences. As such, it may be that Japanese to English translation "predicts" a subject, as one is required in English.

1.2 Politeness Labels

We designed our model to be able to handle both translation and formality classification. While extracting a representation for politeness from the automatically extracted features in a neural network pipeline isn't impossible, that is not what we are trying to achieve. Rather, we would train our model under a supervised learning scheme, where each sentence has a corresponding ground truth translation and formality label attached.

One of the earliest roadblocks we faced is the scarcity of such sentence-formality pairs. Such corpuses are extremely difficult to find in sufficient quantities that would allow for adequate training of a neural classification model. Human annotation is unfortunately not so accessible for Japanese (in terms of pricing) as some other languages. As such, we devised a number of ways to generated such sentence-formality pairs.

1.2.1 Procedural Generation of Politeness Labels

Fortunately for us, this is a topic that had been studied previously by Feely et al. [FHG19], which in turn was based on the Kyoto Text Analysis Toolkit (KyTea) [NNM11]. In their case, Japanese was the destination language; the task was to generate Japanese sentences that would match the formality levels of the input.

The authors identified verb suffixes and copulas as the keys for indentifying sentence formality, with short-form corresponding to informal sentences and long-form corresponding to formal ones. They published a conversion script which would identify and convert any informal verb suffixes into formal verb suffixes, and vice versa. There is a slight complication to this rule, which is that in Japanese there are situations which constrain the verb to be short-form, ie. when it forms part of a verb phrase that isn't the head of the sentence. Fortunately, Japanese is also a language which

observes the SOV order, meaning that the final verb in a sentence is always the head verb and is never constrained to short form. As such, we can first perform sentence tokenisation and then use the long-short form of the final verb/copula in each sentence to determine its formality. There are some exceptions to this SOV structure, especially in the short-form, and especially involving the omission of the verb "desu", (translated as "to be"). Fortunately, in these cases, the final term is an adjective, which in Japanese also conjugate in a way that's distinguishable by politeness.

We thus from their script and make two important modifications. First we adapt their conversion script into a classificaiton script. We also take advantage of the fact that the formality level in a single document should remain the same. This observation can serve multiple purposes:

1 A sanity check for the outputs of our script;

2 Simplify our calculation for classifying sentences that had been segmented into documents.

We performed a sanity check on the legal corpus we mentioned earlier (a corpus entirely made up of informal sentences) and got very positive results: not a single sentence in the corpus of 262000 lines were identified as formal.

Some corpuses are already tokenised at sentence level and do not form part of a document, and we were forced to weaken our common-formality assumption to individual corpus entries (which may contain multiple sentences).

1.2.2 Using Pre-trained Japanese Language Model for Politeness Labelling

While Japanese is not nearly as high-resource as English (GPT-3 was never specifically trained on Japanese, for example), there are nevertheless some available pre-trained Japanese language models that are available. One of the best recent models that were developed is the Japanese BERT trained at Tohoku University. Similar to the original BERT, this model performs a mask-filling task on Japanese sentences.

The obvious way of making use of a pre-trained language model for politeness labelling would be to fine-tuned it on the new task. However, that would require correctly labelled data, leading us to a chicken-and-egg problem. Alternatively we could use the labels that we generated in section 1.2.1, although then we would be constrained by the quality of our previous scoring function.

A slightly modified approach would be: for each sentence, we identify and mask each verb (along with its suffix conjugation) and each copula one at a time. We feed the masked sentences to the language model. We then score each sentence based on whether the language model filled the masks with long forms or short forms.

1.2.2.1 BERT Fine-tuned on Formality Classification The setup for this experiment is simple: we take a pre-trained Japanese language model and attempt to fine tune it with the sentence-formality label pairs that we generated in section 1.2.1. We first try viewing this problem as a scoring problem, with the model tuned to produce a score between 0 and 1 for each document and with a score greater than .5 indicating that the model considers the document to be formal, and vice versa. We tried two different experimental settings, one where ambiguous sentences (those that didn't contain a formality indicator) are mixed in with target score .5, the other where ambiguous sentences are not included.

The other approach is to regard the problem as a two-class classification task. We do not introduce a third class of ambiguous sentences under this setting, since it is questionable from a linguistic point of view to label these sentences as "formality-free".

1.2.2.2 Masking Decisive Verb/Copula This setup doesn't even require any fine-tuning, simply deploying existing mask-fill language models and masking the correct words would suffice.

1.2.3 Other Techniques for Politeness Labelling

Some other techniques for politeness labelling of Japanese and other languages had been proposed, and we will discuss them briefly here. Dugan [Dug20] proposed generating politeness labels from the corresponding English translation, an idea we didn't find convincing due to the inherent problems associated with inferring formality from English which has relatively few clear markers for formality, not the mention the problem with noise introduced by otherwise perfect translations with incorrect formality.

2 Model Architecture

A number of possible architectures could be used to integrate the translation task with the formality labelling task. We implement a number of thems, comparing their performance when their tasks are comparable.

2.1 Base translation model without formality labels

We first build a base translation model that doesn't take formality into account. We will use its performance as a baseline and attempt to beat its performance with our formality-aware model. For test cases where either the source or the target language doesn't have strong formality classes (such as Japanese-English or vice versa), we are only interested in the translation quality. For cases where both the source and the target language have strong formality classes (such as Korean-Japanese), we are also interested in formality class correspondance.

2.2 Base Autoencoder-Reconstructor

This can be seen as a special case of the base translation model, except that the source and the target languages are identical and we measure its ability to reconstruct the original sentence as well as the correct formality label.

2.3 Encoder-Decoder Decomposition

We decompose the usual encoder-decoder, using the encoder output as the basis for the formality labelling task. We build a separate classification module which takes encoder output as input and produces a formality label prediction. During training time, the loss function is constructed as a weighted sum of the translation loss and the formality label loss.

2.4 Formaltiy Autoencoder

This setting can been seen extending the base autoencoder setting. Here we combine the encoder output with the output of the classifier and feed it back to a decoder which emits a sequence in the source language. We then measure how well the model preserves both the meaning and the formality of the original sentence.

2.5 Formality-Aware Translation

This is the generalised version of the formality autoencoder, where the target language is different from the source language (finally the true translation setting). We compare the results with the base translation model without formality labels.

2.6 Formality-Aware Multi-Channel Autoencoder-Reconstructor

2.7 Formality-Aware Multi-Channel Translation

3 Evaluation

We propose a number of different ways to evaluate the translation quality and the sentence reconstruction quality for our different tasks. We analyse the level of correlation between these different metrics on different tasks.

3.1 Alignment BLEU score

We use the NLTK toolkit [BLK09] to evaluate the BLEU score of our translations.

3.2 Sentence Embedding Similarity

Comparing the embedding similarity of two sentences is hardly a trivial problem. In order to make our evaluation task tractable, we use a rather trivial construction where we define the embedding of a sentence as the average embedding of all word embeddings in the sentence. We then define sentence similarity as the dot product between the embedding of the two sentences.

References

[BLK09] Steven Bird, Edward Loper, and Ewan Klein. O'Reilley Media Inc., 2009.

[Dug20] Liam Dugan. "Learning Formality from Japanese-English Parallel Corpora". MA thesis. University of Pennsylvania, 2020. URL: http://liamdugan.com/static/thesis-bb3055b500391857a7b237a22a6f18e5.pdf.

[FHG19] 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*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 45–53. URL: https://www.aclweb.org/anthology/D19-5203.

[NNM11] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. "Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis". In: *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*. Portland, Oregon, USA., 2011. URL: http://www.phontron.com/paper/neubig11aclshort.pdf.