Zero-pronoun Data Augmentation for Japanese-to-English Translation

Ryokan Ri, Toshiaki Nakazawa and Yoshimasa Tsuruoka
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
{li0123, nakazawa, tsuruoka}@logos.t.u-tokyo.ac.jp

Abstract

For Japanese-to-English translation, zero pronouns in Japanese pose a challenge, since the model needs to infer and produce the corresponding pronoun in the target side of the English sentence. However, although fully resolving zero pronouns often needs discourse context, in some cases, the local context within a sentence gives clues to the inference of the zero pronoun. In this study, we propose a data augmentation method that provides additional training signals for the translation model to learn correlations between local context and zero pronouns. We show that the proposed method significantly improves the accuracy of zero pronoun translation with machine translation experiments in the conversational domain.

1 Introduction

While neural machine translation (NMT) has demonstrated high performance in single-sentence translation, it is still challenging to handle linguistic phenomena involving discourse contexts. One such issue is the translation of *zero pronouns* (ZP) in Japanese-to-English translation. In Japanese, subjects and objects are often omitted when the listener can infer them from the context. However, when translating them into English, the omitted words must be explicitly translated in most cases. For example, in the following sentence, the subject omitted in Japanese is the first person, and *I* has to be output in English.

うなぎが　食べたいな
unagi-ga　tabe-tai-na
eel-OBJ　eat-want-PARTICLE
I feel like eating eel.

The prediction of ZPs, essentially, requires understanding the topic and old information in the discourse, or referring to the world knowledge. On the other hand, linguistic information within the sentence may provide some clues (Kudo et al., 2015). For example, in the sentence above, the auxiliary verb たい (*want*) suggests that the sentence expresses a subjective statement and thus the missing pronoun is the first person. Here we refer to such information as *local context*.

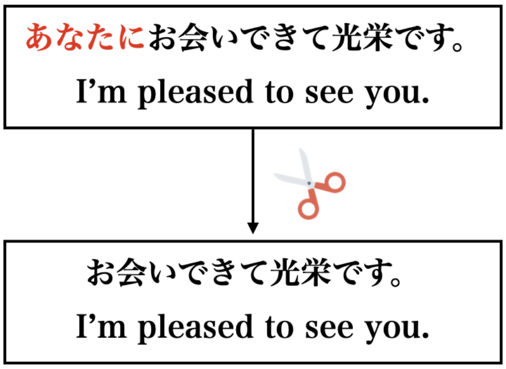

Figure 1: The proposed method: ZP data augmentation

Correlations between local context and ZPs can be learned by the standard single-sentence neural machine translation, but it may not be possible under low-resource conditions. For example, the translation of conversations, which usually contain a large number of ZPs, is currently one of the under-resourced domains.

To address this problem, we propose **zero pronoun data augmentation** to facilitate learning correlations between local context and ZPs (Figure 1). We augment the training data by deleting personal pronouns in the source Japanese sentence. This creates parallel data that include ZPs and provides additional training signals to learn to predict ZPs. Our method is simple yet effective: it does not require any modification to the model architecture nor additional computation at inference time, but significantly improves the accuracy of the ZP translation.

Proceedings of the 8th Workshop on Asian Translation, pages 117–123
Bangkok, Thailand (online), August 5-6, 2021. ©2021 Association for Computational Linguistics

2 Related Work

2.1 Contextual Neural Machine Translation

As the quality of single-sentence machine translation has improved dramatically with the advent of neural machine translation (Sutskever et al., 2014; Vaswani et al., 2017), translation models that take wider contexts into account have seen a surge of interest (Jean et al., 2017; Bawden et al., 2018; Voita et al., 2019b,a; Ma et al., 2020; Saunders et al., 2020). In contrast to the studies trying to incorporate information outside the sentence, in this work, we propose a method to improve zero-pronoun translation by only considering the information within the sentence, but we also explore the effect of combining our method with a contextual machine translation model.

2.2 ZP Resolution in Japanese

In some languages, pronouns are sometimes omitted when they are inferable from the context. Such languages are called pro-drop languages and the omitted pronouns are called ZPs.

The translation of ZPs poses a challenge when the corresponding pronoun is syntactically required on the target language side: the model has to infer the omitted pronoun. The task of identifying the omitted pronouns is called ZP resolution and for Japanese, this has been a long-standing problem (Isozaki and Hirao, 2003; Sasano et al., 2008; Imamura et al., 2009; Shibata and Kurohashi, 2018). Japanese is one of the most difficult languages because Japanese words usually do not have any inflectional forms that depend on the omitted pronoun, unlike other pro-drop languages such as Portuguese and Spanish in which ZPs can be inferred from the grammatical case of other words.

Still, Japanese sentences sometimes contain expressions indicative of the missing pronoun. For example, Japanese honorifics naturally indicate the subject is the second person. In this work, we do not explicitly solve ZP resolution but let the translation model learn heuristic relations between ZPs and local context within the sentence (Hangyo et al., 2013; Kudo et al., 2015) and produce appropriate English pronouns.

2.3 ZPs in Translation

In the context of statistical machine translation, Japanese ZPs are explicitly predicted by considering verbal semantic attributes (Nakaiwa and Ikehara, 1992), local context in the source and target sentence (Kudo et al., 2015), and incorporated into the resulting translation.

On the other hand, in neural machine translation, the missing pronouns can be automatically inferred by the translation model because of the nature of end-to-end learning, although the correctness cannot be guaranteed. To improve the quality of ZP translation, previous studies have explored a multi-task approach with ZP prediction (Wang et al., 2016, 2019).

In this study, we propose a ZP data augmentation method to provide additional training signals useful to correctly translate ZPs.

3 Is Local Context Useful for Predicting Zero Pronouns?

Our proposed method is based on the assumption that local context in Japanese sentences is useful for predicting ZPs. We begin by analyzing to what extent ZPs can be inferred from local context, and what kind of local context is useful.

For the analysis, we use the Business Scene Dialogue Corpus (Rikters et al., 2019), which is a Japanese and English parallel corpus in the conversational domain. Besides the published data, we also use the in-house version of the corpus, which amounts to a total of 104,961 sentence pairs.

3.1 Identifying sentence pairs that contain ZPs.

As the corpus does not contain annotations of ZPs, we first identify sentence pairs that contain zero pronouns. We exploit the word alignment information from parallel sentences to detect ZPs. The specific procedure is as follows.

1. We obtain the word alignments of the parallel data with `GIZA++`[1]. We use `Mecab`[2] for Japanese word segmentation, `spaCy`[3] for English.

2. When a pronoun in an English sentence is associated with `NULL`, the pronoun in the English sentence is considered to correspond to a ZP in the Japanese sentence.

The resulting number of pronouns is shown in Figure 2. It can be seen that in the conversational domain, the first person pronoun *I* and the second

[1] https://github.com/moses-smt/giza-pp

[2] https://taku910.github.io/mecab/

[3] https://spacy.io/

	I	you	we	they	he	she	us	them	him	her
baseline	35.9	25.4	11.0	3.7	2.2	0.0	2.2	1.9	1.2	0.9
logistic regression	78.2	46.3	17.3	3.8	3.1	0.0	3.6	0.2	0.2	2.9

Table 1: Recall scores of ZP predictions for each pronoun.

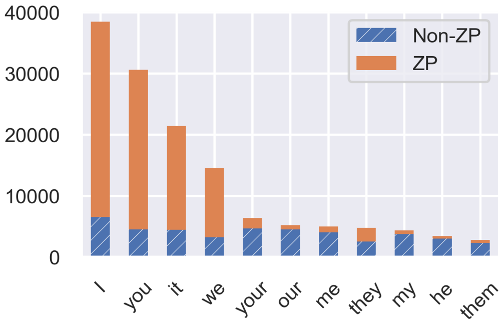

Figure 2: The number of English pronouns in the analyzed data. ZP stands for those whose corresponding pronoun does not appear in the Japanese text.

person pronoun *you* occur frequently and most of them (80% ∼) are omitted in Japanese. More infrequent pronouns are less likely to be ZPs.

3.2 Extracting local context that co-occurs with ZPs

To associate the detected ZPs with local context in Japanese sentences, we extract the words that appear in their predicates. We did not use a Japanese syntactic analyzer to detect ZPs but they are associated with the English pronouns by alignment. Therefore, we decided to exploit the alignment information to extract the predicates. We extract the predicates of the English pronoun and the corresponding words in the Japanese sentence. Specifically, the following steps were taken.

3. We obtain the dependency tree of the English sentence with `spaCy` and extract the pronoun's head.

4. The Japanese word aligned to the pronoun's head and its subsequent functional words [4] are extracted as local context.

3.3 Predicting ZPs from Local Context

To investigate the extent to which ZPs can be predicted from local context, we conducted an analysis by training a logistic regression classifier [5]. The classifier takes the unigrams, bi-grams, and trigrams extracted from local context in the Japanese sentence and predicts the associated pronoun in the English sentence.

The recall scores of each pronoun obtained with five-fold cross-validation are shown in Table 1. As a baseline, we adopt the score of random prediction according to the training distribution of pronouns.

One can see that the frequent pronouns such as *I, you, we* can be predicted with significantly higher accuracy than the baseline when local context is used (around 6 to 43 points of improvement). In contrast, the other infrequent pronouns display similar or lower values compared to the baseline. In summary, we can see that local context is predictive of the frequent pronouns but not for the infrequent ones.

To investigate what kind of local context is useful for prediction, for each output label (*i.e.*, pronoun) of the logistic regression classifier, we extracted the input features with higher values in the corresponding weights. As a result, the following words are interpreted to be relevant.

The first person singular *I* verbs related to recognition (思う (think), わかる (understand), 感じる (feel)); humble words (申し上げる、存る); and auxiliary verbs expressing desire (たい).

The second person singular *you* suffixes expressing questions (かな？, ました？); speculations (でしょ, だろ？), honorifics (仰る, いただける).

The first person plural *we* obligations (なきゃ, べき), desire (たい).

For the other pronouns, no local contexts were found to be interpretable as useful for prediction.

4 ZP Data Augmentation

In the previous section, we confirmed that local context is useful for predicting ZPs. In this section, we examine the usefulness of ZP data augmentation for machine translation.

[4] In this case, the function words are defined as words with one of the following part of speeches defined in `Mecab`: ["particle", "auxiliary verb", "symbol"].

[5] We use the implementation of the `scikit-learn` library with the default hyperparameters.

	1to1	2to1
baseline	17.07±0.16 / 83.6±1.1	17.07±0.26 / 89.36±0.9
baseline+pro_aug	17.07±0.19 / 92.32±1.8	17.11±0.23 / 92.17±1.1

Table 2: Evaluation of the model with ZP data augmentation. The scores on the table are BLEU / ZP evaluation accuracy. The mean and standard deviation of five runs with different random seeds are reported.

The method artificially creates training data containing ZPs by deleting pronouns in the source Japanese sentence along with the following particles. The pronouns to be deleted are detected by string matching with manually created lists (Appendix A). The augmented data is supposed to provide useful training signals for learning correlations between ZPs and local context.

4.1 Experimental Setups

Corpus We use the Document-aligned Japanese-English Conversation Parallel Corpus (Rikters et al., 2020). We also add an in-house conversational parallel corpus to the training data. The statistics of the corpus are shown in Table 3.

train	train+pro_aug	dev	test
246,541	282,952	2,051	2,020

Table 3: The number of sentences in the corpus.

Model Transformer (Vaswani et al., 2017) was used as the translation model. We adopt the hyperparameters recommended for the corpus of our size in Araabi and Monz (2020) (Appendix B). In addition to the single-sentence translation, we also experimented with the 2to1 setting (Tiedemann and Scherrer, 2017), in which the previous sentence in the document is added to the input.

Evaluation We evaluate the overall translation quality on the test set with BLEU (Papineni et al., 2002). We also conduct a targeted evaluation with the ZP evaluation dataset for Japanese-to-English translation (Shimazu et al., 2020). The ZP evaluation dataset contains 724 triples of a source sentence, a target sentence with a correct pronoun, and one with an incorrect pronoun. To evaluate a translation model, we see if the model assigns a lower perplexity to the correct target sentence, and calculate the accuracy.

4.2 Results

The results of the experiment are shown in Table 2. We can observe that ZP data augmentation does not improve the BLEU score, but significantly improves the accuracy of ZP evaluation in both the 1to1 (83.6% to 92.3%) and 2to1 settings (89.3% to 92.1%). Our method yields a similar degree of improvement to the 2to1 setting in the ZP evaluation without any computational overhead at the inference time.

We also confirm that adding the previous context (2to1) does not improve BLEU but pronoun translation (83.6% to 89.3%), which conforms to observations in the previous study (Jean et al., 2017; Shimazu et al., 2020). However, this is not the case with the ZP data augmentation (92.3% to 92.1%). We speculate that this is because longer inputs in the 2to1 setting make it more difficult for the model to find correlations between ZPs and local context.

5 Conclusion

To address the problem of zero pronoun translation, we proposed zero pronoun data augmentation. Through the analysis with the Japanese-English conversational parallel corpus, we showed that zero pronouns in Japanese sentences can be predicted to some extent from local context within the sentence. In the conversational translation experiment, we compared a translation model trained on the augmented data with the baseline and demonstrate that our method significantly improves the accuracy of zero pronoun translation.

Nevertheless, zero pronoun data augmentation does not solve the cases where the information necessary for zero pronoun translation exists outside the sentence. Also, the analysis suggests that local context is useful for predicting frequent pronouns such as the first and second-person pronouns, but not for the third-person pronouns. An interesting avenue for future work is to explicitly incorporate discourse-level contextual information such as topics or people involved in the conversation into the translation models.

References

Ali Araabi and Christof Monz. 2020. Optimizing Transformer for Low-Resource Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2013. Japanese Zero Reference Resolution Considering Exophora and Author/Reader Mentions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Kenji Imamura, Kuniko Saito, and Tomoko Izumi. 2009. Discriminative Approach to Predicate-Argument Structure Analysis with Zero-Anaphora Resolution. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*.

Hideki Isozaki and Tsutomu Hirao. 2003. Japanese Zero Pronoun Resolution based on Ranking Rules and Machine Learning. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does Neural Machine Translation Benefit from Larger Context? *ArXiv*, abs/1704.05135.

Taku Kudo, Hiroshi Ichikawa, and Hideto Kazawa. 2015. Language independent null subject prediction for statistical machine translation. In *Proceedings of the Twenty-first Annual Meeting of the Association for Natural Language Processing*.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A Simple and Effective Unified Encoder for Document-Level Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Hiromi Nakaiwa and Satoru Ikehara. 1992. Zero Pronoun Resolution in a Machine Translation System by using Japanese to English Verbal Semantic Attributes. In *Third Conference on Applied Natural Language Processing*.

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*.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2019. Designing the Business Conversation Corpus. In *Proceedings of the 6th Workshop on Asian Translation*.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2020. Document-aligned Japanese-English Conversation Parallel Corpus. In *Proceedings of the Fifth Conference on Machine Translation*.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2008. A Fully-Lexicalized Probabilistic Model for Japanese Zero Anaphora Resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics*.

Danielle Saunders, Felix Stahlberg, and Bill Byrne. 2020. Using Context in Neural Machine Translation Training Objectives. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Tomohide Shibata and Sadao Kurohashi. 2018. Entity-Centric Joint Modeling of Japanese Coreference Resolution and Predicate Argument Structure Analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Sho Shimazu, Sho Takase, Toshiaki Nakazawa, and Naoaki Okazaki. 2020. Evaluation Dataset for Zero Pronoun in Japanese to English Translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, volume 27.

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-Aware Monolingual Repair for Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Longyue Wang, Zhaopeng Tu, Xing Wang, and Shuming Shi. 2019. One Model to Learn Both: Zero Pronoun Prediction and Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016. A Novel Approach to Dropped Pronoun Translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

A The pronoun and particle list for pronoun data augmentation

The deletion of pronouns was done by enumerating all combinations from the list of pronouns (Table 4) and particles (Table 5) and deleting strings that correspond to the pattern from the sentence.

First person singular	私, わたし, 僕, ぼく, 俺, おれ, わたくし, オレ, ウチ
First person plural	我々, 僕ら, われわれ, 僕達, 僕たち, 私達
Second person singular	貴方, 貴女, あなた, お前, おまえ, 君, あんた
First person plural	君たち, みなさま
Third person singular	彼, 彼女, あいつ
Third person plural	彼ら, 彼女ら, みんな, 皆, 皆んな, みなさん, 奴ら

Table 4: The list of pronouns for pronoun deletion

Nominative	は, が
Accusative	を
Dative	に
Possessive	の
Others	も, の方から, のほうから, の方に, のほうに, の方で のこと, の事, のほうで, から, 、

Table 5: The list of particles for pronoun deletion

B Hyperparameters for the Machine Translation Experiment

We choose the hyperparameters of the Transformer model recommended in (Araabi and Monz, 2020).

layers	5
model size	512
feed-forward dimension	2048
number of attention heads	4
encoder/decoder layer dropout	0/0.1
src/tgt word dropout	0.2/0.2
label_smoothing	0.3
optimizer	Adam with the Noam Learning rate schedule
warmup steps	8000

Table 6: Hyperparameters for the Transformer model.