Japanese Zero Anaphora Resolution Can Benefit from Parallel Texts Through Neural Transfer Learning

Masato Umakoshi, Yugo Murawaki, Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{umakoshi, murawaki, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Parallel texts of Japanese and a non-pro-drop language have the potential of improving the performance of Japanese zero anaphora resolution (ZAR) because pronouns dropped in the former are usually mentioned explicitly in the latter. However, rule-based cross-lingual transfer is hampered by error propagation in an NLP pipeline and the frequent lack of transparency in translation correspondences. In this paper, we propose implicit transfer by injecting machine translation (MT) as an intermediate task between pretraining and ZAR. We employ a pretrained BERT model to initialize the encoder part of the encoder-decoder model for MT, and eject the encoder part for fine-tuning on ZAR. The proposed framework empirically demonstrates that ZAR performance can be improved by transfer learning from MT. In addition, we find that the incorporation of the masked language model training into MT leads to further gains.

1 Introduction

Figuring out who did what to whom is an essential part of natural language understanding. This is, however, especially challenging for so-called pro-drop languages like Japanese and Chinese because they usually omit pronouns that are inferable from context. The task of identifying the referent of such a dropped element, as illustrated in Figure 1(a), is referred to as *zero anaphora resolution* (ZAR). Although Japanese ZAR saw a performance boost with the introduction of BERT (Ueda et al., 2020; Konno et al., 2020), there is still a good amount of room for improvement.

A major barrier to improvement is the scarcity of training data. The number of annotated sentences is the order of tens of thousands or less (Kawahara et al., 2002; Hangyo et al., 2012; Iida et al., 2017), and the considerable linguistic expertise required for annotation makes drastic corpus expansion impractical.

Previous attempts to overcome this limitation exploit orders-of-magnitude larger parallel texts of Japanese and English, a non-pro-drop language (Nakaiwa, 1999; Furukawa et al., 2017). The key idea is that Japanese zero pronouns can be recovered from parallel texts because they are usually mentioned explicitly in English, as in Figure 1(b). If translation correspondences are identified and the anaphoric relation in English is identified, then we can identify the antecedent of the omitted argument in Japanese.

Their rule-based transfer from English to Japanese had met with limited success, however. It is prone to error propagation due to its dependence on word alignment, parsing, and English coreference resolution. More importantly, the great linguistic differences between the two language often lead to parallel sentences without transparent syntactic correspondences (Figure 1(c)).

In this paper, we propose neural transfer learning from machine translation (MT). By generating English translations, a neural MT model should be able to implicitly recover omitted Japanese pronouns, thanks to its expressiveness and large training data. We expect the knowledge gained during MT training to be transferred to ZAR. Given that state-of-the-art ZAR models are based on BERT (Ueda et al., 2020; Konno et al., 2020, 2021), it is a natural choice to explore *intermediate task* transfer learning (Phang et al., 2018; Wang et al., 2019a; Pruksachatkun et al., 2020; Vu et al., 2020): A pretrained BERT model is first trained on MT and the resultant model is then fine-tuned on ZAR.[1]

A key challenge to this approach is a mismatch in model architectures. While BERT is an encoder, the dominant paradigm of neural MT is the encoder-decoder. Although both share Transformer (Vaswani et al., 2017) as the building block,

[1] In our preliminary experiments, we tested encoder-decoder pretrained models with no success. We briefly revisit this in Section 4.7.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1920–1934
November 7–11, 2021. ©2021 Association for Computational Linguistics

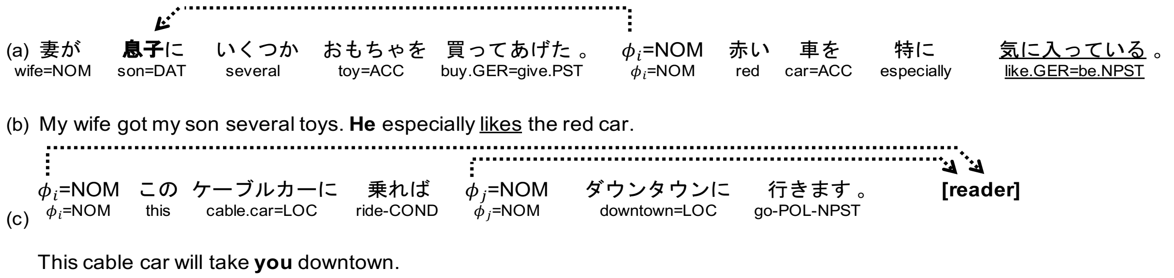

(a) 妻が 息子に いくつか おもちゃを 買ってあげた。 ϕi=NOM 赤い 車を 特に 気に入っている。
wife=NOM son=DAT several toy=ACC buy.GER=give.PST ϕi=NOM red car=ACC especially like.GER=be.NPST
(b) My wife got my son several toys. He especially likes the red car.
(c) ϕi=NOM この ケーブルカーに 乗れば ϕj=NOM ダウンタウンに 行きます。 [reader]
ϕi=NOM this cable.car=LOC ride-COND ϕj=NOM downtown=LOC go-POL-NPST
This cable car will take you downtown.

Figure 1: (a) An example of Japanese zero anaphora. The nominative argument of the underlined predicate is omitted. The goal of the task is to detect the omission and to identify its antecedent "son". (b) The corresponding English text. The omitted argument in Japanese is present as a pronoun in English. (c) A Japanese-English pair (Nabeshima and Brooks, 2020, p. 74) whose correspondences are too obscure for rule-based transfer. Because Japanese generally avoids having inanimate agents with animate patients, the English inanimate-subject sentence corresponds to two animate-subject clauses in Japanese, with two exophoric references to the reader (i.e., you).

it is non-trivial to combine the two distinct architectures, with the goal to help the former.

We use a pretrained BERT model to initialize the encoder part of the encoder-decoder model for MT. While this technique was previously used by Imamura and Sumita (2019) and Clinchant et al. (2019), they both aimed at improving MT performance. We show that by ejecting the encoder part for use in fine-tuning (Figure 2), we can achieve performance improvements in ZAR. We also demonstrate further improvements can be brought by incorporating encoder-side masked language model (MLM) training into the intermediate training on MT.

2 Related Work

2.1 Zero Anaphora Resolution (ZAR)

ZAR has been extensively studied in major East Asian languages, Chinese and Korean as well as Japanese, which not only omit contextually inferable pronouns but also show no verbal agreement for person, number, or gender (Park et al., 2015; Yin et al., 2017; Song et al., 2020; Kim et al., 2021). While supervised learning is the standard approach to ZAR (Iida et al., 2016; Ouchi et al., 2017; Shibata and Kurohashi, 2018), training data are so small that additional resources are clearly needed. Early studies work on case frame construction from a large raw corpus (Sasano et al., 2008; Sasano and Kurohashi, 2011; Yamashiro et al., 2018), pseudo training data generation (Liu et al., 2017), and adversarial training (Kurita et al., 2018). These efforts are, however, overshadowed by the surprising effectiveness of BERT's pretraining (Ueda et al., 2020; Konno et al., 2020).

Adopting BERT, recent studies seek gains through multi-task learning (Ueda et al., 2020), data augmentation (Konno et al., 2020), and an intermediate task tailored to ZAR (Konno et al., 2021). The multi-task learning approach of Ueda et al. (2020) covers verbal predicate analysis (which subsumes ZAR), and nominal predicate analysis, coreference resolution, and bridging anaphora resolution. Their method is used as a state-of-the-art baseline in our experiments.

Konno et al. (2020) perform data augmentation by simply masking some tokens. They found that performance gains were achieved by selecting target tokens by part of speech. Konno et al. (2021) introduce a more elaborate masking strategy as a ZAR-specific intermediate task They spot multiple occurrences of the same noun phrase, mask one of them, and force the model to identify the pseudo-antecedent.

Our use of parallel texts in ZAR is inspired by Nakaiwa (1999) and Furukawa et al. (2017), who identify a multi-hop link from a Japanese zero pronoun to its Japanese antecedent via English counterparts. Their rule-based methods suffer from accumulated errors and syntactically non-transparent correspondences. In addition, they do not handle *inter-sentential* anaphora, a non-negligible subtype of anaphora we cover in this paper.

While we exploit MT to improve the performance of ZAR, the exploitation in the reverse direction has been studied. A line of research has been done on Chinese *zero pronoun prediction* (ZPP) with a primary aim of improving Chinese-English translation (Wang et al., 2016, 2018, 2019b). ZPP is different from ZAR in that it does not identify antecedents. This is understandable given that classification of zero pronouns into overt ones suffices for MT. Although Wang et al. (2019b) report mutual benefits between MT and ZPP, it remains an

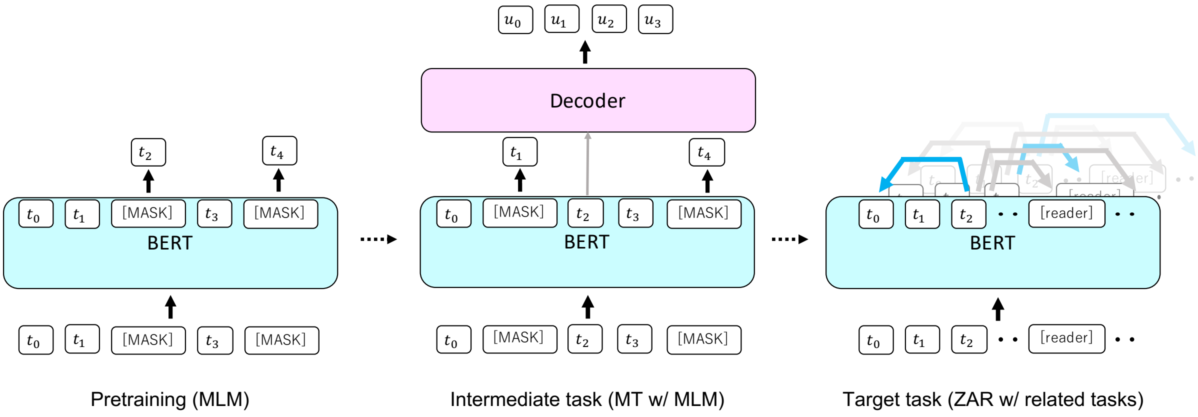

u_0 u_1 u_2 u_3
Decoder
t_2 t_4
t_1 t_4
t_0 t_1 [MASK] t_3 [MASK]
BERT
t_0 [MASK] t_2 t_3 [MASK]
t_0 t_1 t_2 [reader]
Pretraining (MLM)
Intermediate task (MT w/ MLM)
Target task (ZAR w/ related tasks)

Figure 2: Overview of the proposed method. **Left**: The model is pretrained with the masked language model (MLM) objective (known as BERT). **Center**: The pretrained BERT is used to initialize the encoder part of the encoder-decoder, which is trained on MT with the MLM objective. **Right**: The encoder is extracted from the MT model and is fine-tuned on ZAR and related tasks. Note that some special tokens are omitted for simplicity.

open question whether MT helps ZAR as well.

2.2 MT as an Intermediate Task

Inspired by the great success of the pretraining/fine-tuning paradigm on a broad range of tasks (Peters et al., 2018; Devlin et al., 2019), a line of research inserts an intermediate task between pretraining and fine-tuning on a target task (Phang et al., 2018; Wang et al., 2019a; Pruksachatkun et al., 2020). However, Wang et al. (2019a) found that MT used as an intermediate task led to performance degeneration in various target tasks, such as natural language inference and sentiment classification.[2] They argue that the considerable difference between MLM pretraining and MT causes *catastrophic forgetting* (CF). Pruksachatkun et al. (2020) suggest injecting the MLM objective during intermediate training as a possible way to mitigate CF, which we empirically test in this paper.

2.3 Use of BERT in MT

Motivated by BERT's success in a wide range of applications, some studies incorporate BERT into MT models. A straightforward way to do this is to initialize the encoder part of the encoder-decoder with pretrained BERT, but it has had mixed success at best (Clinchant et al., 2019; Zhu et al., 2020).

Abandoning this approach, Zhang et al. (2020) simply use BERT as a supplier of context-aware embeddings to their own encoder-decoder model. Similarly, Guo et al. (2020) stack adapter layers on top of two frozen BERT models to use them as the encoder and decoder of a non-autoregressive MT model. However, these methods cannot be adopted for our purpose because we want BERT itself to learn from MT.

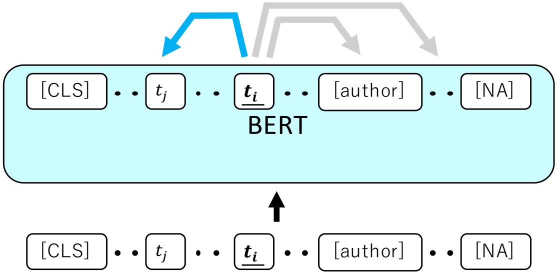


Figure 3: ZAR as argument selection.

Imamura and Sumita (2019) manage to maintain the straightforward approach by adopting a two-stage training procedure: In the first stage, only the decoder is updated with the encoder frozen, while in the second stage, the entire model is updated. Although they offer some insights, it remains unclear how best to exploit BERT when MT is an intermediate task, not the target task.

3 Proposed Method

We adopt a ZAR model of Ueda et al. (2020), which adds a thin layer on top of BERT during fine-tuning to solve ZAR and related tasks (Section 3.1). Instead of directly moving from MLM pretraining to fine-tuning on ZAR, we inject MT as an intermediate task (Section 3.2). In addition, we introduce the MLM training objective during the intermediate training (Section 3.3).

[2] We suspect that the poor performance resulted in part from their excessively simple decoder, a single-layer LSTM.

3.1 BERT-based Model for ZAR

ZAR as argument selection As illustrated in Figure 3, the basic idea behind BERT-based ZAR is that given the powerful neural encoder, the joint task of omission detection and antecedent identification can be formalized as argument selection (Shibata and Kurohashi, 2018; Kurita et al., 2018; Ueda et al., 2020). Omission detection concerns whether a given predicate has an argument for a given case (relation). If not, the model must point to the special token `[NULL]`. Otherwise the model must identify the antecedent of the zero pronoun by pointing either to a token in the given text or to a special token reserved for exophora. Note that by getting the entire document as the input, the model can handle inter-sentential anaphora as well as intra-sentential anaphora. In practice, the input length limitation of BERT forces us to implement a sliding window approach. Also note that in this formulation, ZAR is naturally subsumed into verbal predicate analysis (VPA), which also covers instances where the predicate and the argument have a dependency relation and only the case marker is absent.

Formally, the probability of the token t_j being the argument of the predicate t_i for case c is:

$$P(t_j|t_i, c) = \frac{\exp(s_c(t_j, t_i))}{\sum_{j'} \exp(s_c(t_{j'}, t_i))} \quad (1)$$

$$s_c(t_j, t_i) = \boldsymbol{v}^\mathsf{T} \tanh(W_c \boldsymbol{t}_j + U_c \boldsymbol{t}_i) \quad (2)$$

where $\boldsymbol{t}_i$ is the context-aware embedding of t_i provided by BERT, W_c and U_c are case-specific weight matrices, and $\boldsymbol{v}$ is a weight vector shared among cases. We output t_j with the highest probability. For each predicate, we repeat this for the nominative (NOM), accusative (ACC), and dative (DAT) cases, and another nominative case for the double nominative construction (NOM2).

Input representations We append some special tokens at the end of the input sequence: `[NULL]` for null arguments, and `[author]`, `[reader]`, and `[unspecified person]` for exophora. The special token `[NA]` is also supplied for the reason given in the next paragraph. As is usual for BERT, the special tokens `[CLS]` and `[SEP]` are inserted at the beginning and end of the sequence, respectively. If a predicate or argument candidate is split into two or more subwords, the initial subword is used for argument selection.

Multi-task learning Following Ueda et al. (2020), we use a single model to simultaneously perform verbal predicate analysis (VPA), nominal predicate analysis (NPA), bridging anaphora resolution (BAR), and coreference resolution (CR). NPA is a variant of VPA in which verb-like nouns serve as predicates taking arguments. BAR is a special kind of anaphora resolution in which the antecedent fills a semantic gap of the anaphor (e.g., "price" takes something priced as its argument). CR identifies the antecedent and anaphor that refer to the same real-world entity, with the special token `[NA]` reserved for nouns without coreferent mentions. All of the four tasks can be formalized as argument selection as in Eq. (1). By sharing the BERT encoder, these interrelated tasks have an influence on each other during training. In addition, case-specific weights are shared between VPA and NPA while separate weights are used for BAR and CR. During training, we compute the losses equally for the four tasks.

3.2 MT as an Intermediate Task

Our main proposal is to use MT as an intermediate task prior to fine-tuning on ZAR. Following Imamura and Sumita (2019) and Clinchant et al. (2019), we use a pretrained BERT to initialize the encoder part of the Transformer-based encoder-decoder model while the decoder is randomly initialized. After the intermediate training on MT, we extract the encoder and move on to fine-tuning on ZAR and related tasks (Figure 2).

Specifically, we test the following two procedures for intermediate training:

One-stage optimization The entire model is updated throughout the training.

Two-stage optimization In the first stage, the encoder is frozen and only the decoder is updated. In the second stage, the entire model is updated (Imamura and Sumita, 2019).

3.3 Incorporating MLM into MT

As discussed in Section 2.2, MT as an intermediate task reportedly harms target-task performance, probably because MT forces the model to forget what it has learned from MLM pretraining (catastrophic forgetting). To overcome this problem, we incorporate the MLM training objective into MT, as suggested by Pruksachatkun et al. (2020). Specifically, we mask some input tokens on the encoder

	Web	News
# of sentences	16,038	11,276
# of zeros	30,852	27,062

Table 1: The numbers of sentences and zero anaphors in each corpus.

side and force the model to recover the original tokens, as depicted in the center of Figure 2.

Our masking strategy is the same as BERT's (Devlin et al., 2019): We choose 15% of the tokens at random and 80% of them are replaced with `[MASK]`, 10% of them with a random token, and the rest are unchanged. The corresponding losses are added to the MT loss function.

4 Experiments

4.1 Datasets

ZAR We used two corpora in our experiments: the Kyoto University Web Document Lead Corpus (Hangyo et al., 2012) and the Kyoto University Text Corpus (Kawahara et al., 2002). Based on their genres, we refer to them as the **Web** and **News**, respectively. These corpora have been widely used in previous studies (Shibata and Kurohashi, 2018; Kurita et al., 2018; Ueda et al., 2020). They contained manual annotation for predicate-argument structures (including zero anaphora) as well as word segmentation, part-of-speech tags, dependency relations, and coreference chains. We split the datasets into training, validation, and test sets following the published setting, where the ratio was around 0.75:0.1:0.15. Key statistics are shown in Table 1.

MT We used a Japanese-English parallel corpus of newspaper articles distributed by the Yomiuri Shimbun.[3] It consisted of about 1.3 million sentence pairs[4] with sentence alignment scores. We discarded pairs with scores of 0. Because the task of interest, ZAR, required *inter-sentential* reasoning, consecutive sentences were concatenated into chunks, with the maximum number of tokens equal to that of ZAR. As a result, we obtained around 373,000, 21,000, and 21,000 chunks for the training, validation, and test data, respectively.

Japanese sentences were split into words using the morphological analyzer MeCab with the Juman dictionary (Kudo et al., 2004).[5] Both Japanese and English texts underwent subword tokenization. We used Subword-NMT (Sennrich et al., 2016) for Japanese and SentencePiece (Kudo and Richardson, 2018) for English. We used separate vocabularies for Japanese and English, with the vocabulary sizes of around 32,000 and 16,000, respectively.

4.2 Model Settings

BERT We employed a Japanese BERT model with BPE segmentation distributed by NICT.[6] It had the same architecture as Google's BERT-Base (Devlin et al., 2019): 12 layers, 768 hidden units, and 12 attention heads. It was trained on the full text of Japanese Wikipedia for approximately 1 million steps.

MT We used the Transformer encoder-decoder architecture (Vaswani et al., 2017). The encoder was initialized with BERT while the decoder was a randomly initialized six-layer Transformer. The numbers of hidden units and heads were set to be the same as BERT's (i.e., 768 units and 12 attention heads). We adopted Adam (Kingma and Ba, 2017) as the optimizer. We set the total number of epochs to 50. In two-stage optimization, the encoder was frozen during the first 15 epochs, then the entire model was updated for the remaining 35 epochs. We set a mini-batch size to about 500. The details of hyper-parameters are given in Appendix A.

ZAR For a fair comparison with Ueda et al. (2020), we used almost the same configuration as theirs. We dealt with all subtypes of ZAR: intra-sentential anaphora, inter-sentential anaphora, and exophora. For exophora, we targeted `[author]`, `[reader]`, and `[unspecified person]`. We set the maximum sequence length to 128.[7] All documents from the Web met this limitation. In the News corpus, however, many documents exceeded the sequence length of 128. For such documents, we divided the document into multiple parts such that it had the longest preceding contexts. The evaluation of ZAR was relaxed using a gold coreference chain. The model was trained on the mixture of both corpora and evaluated on each corpus. We used almost the same

[3] https://database.yomiuri.co.jp/about/glossary/

[4] We counted the Japanese sentences since there were one-to-many mappings.

[5] 9 pairs for which morphological analysis failed were removed.

[6] https://alaginrc.nict.go.jp/nict-bert/index.html

[7] We tested longer maximum sequence lengths (256 and 512) but ended up with poorer performance.

Models	Web	News
Shibata and Kurohashi (2018)	58.1	35.6
Kurita et al. (2018)[8]	58.4	-
Ueda et al. (2020)[9]	70.3	56.7
+MT	71.4*	57.7
+MT w/ MLM	**71.9**	**58.3**

Table 2: F1 scores on the test sets. *: with two-stage optimization.

hyper-parameters as Ueda et al. (2020), which are included in Appendix B.

We decided to tune the training epochs for MT since we found that it slightly affected ZAR performance. We collected checkpoints at the interval of 5 epochs out of 45 epochs, in addition to the one with the lowest validation loss. They were all trained on ZAR, and we chose the one with the highest score on the validation set. We ran the model with 3 seeds on MT and with 3 seeds on ZAR, which resulted in 9 seed combinations. We report the mean and the standard deviation of the 9 runs.

4.3 Results

Table 2 summarizes the experimental results. Our baseline is Ueda et al. (2020), who drastically outperformed previous models, thanks to BERT. **+MT** refers to the model with intermediate training on MT while **+MT w/ MLM** corresponds to the model that incorporated the MLM objective into MT. We can see that MT combined with MLM performed the best and that the gains reached 1.6 points for both the Web and News.

Tables 3 and 4 provide more detailed results. For comparison, we performed additional pretraining with ordinary MLM on the Japanese part of the parallel corpus (denoted as **+MLM**), because the possibility remained that the model simply took advantage of additional data. The subsequent two blocks compare **one-stage** (unmarked) optimization with **two-stage** optimization. MT yielded gains on all settings. The gains were consistent across anaphora categories. Although **+MLM** somehow beat the baseline, it was outperformed by most models trained on MT, ruling out the possibility that the gains were solely attributed to extra data. We can conclude that Japanese ZAR benefits from parallel texts through neural transfer learning.

Two-stage optimization showed mixed results. It worked for the Web but did not for the News. What is worse, its combination with MLM led to performance degeneration on both datasets.

MLM achieved superior performance as it worked well in all settings. The gains were larger with one-stage optimization than with two-stage optimization (1.4 vs. 0.3 on the Web).

4.4 Translation of Zero Pronouns

The experimental results demonstrate that MT helps ZAR, but why does it work? Unfortunately, conventional evaluation metrics for MT (e.g., BLEU) reveal little about the model's ability to handle zero anaphora. To address this problem, Shimazu et al. (2020) and Nagata and Morishita (2020) constructed Japanese-English parallel datasets that were designed to automatically evaluate MT models with regard to the translation of Japanese zero pronouns (ZPT). We used Shimazu et al.'s dataset for its larger data size.[10]

To facilitate automatic evaluation of ZPT, this dataset paired a correct English sentence with an incorrect one. All we had to do was to calculate the ratio of instances for which the model assigned higher translation scores to the correct candidates. The only difference between the two sentences involved the translation of a Japanese zero pronoun. To choose the correct one, the MT model must sometimes refer to preceding sentences. As in intermediate training, multiple source sentences were fed to the model to generate multiple target sentences. We prepended as many preceding sentences as possible given the limit of 128 tokens.

In addition, this dataset recorded d, the sentence-level distance between the zero pronoun in question and its antecedent. The number of instances with $d = 0$ was 218 while the number of remaining instances was 506. We regarded the former as the instances of *intra-sentential* anaphora and the latter as the instances of *inter-sentential* anaphora.

We chose the model with the best performance (i.e., one-stage optimization with MLM). For each checkpoint we collected during intermediate training, we (1) measured the ZPT accuracy and (2) fine-tuned it to obtain the F1 score for ZAR. As before,

[8] Not a strict comparison since Kurita et al. (2018) ignored inter-sentential anaphora.

[9] We refer to errata posted on the first author's website: https://nobu-g.github.io/pub/COLING2020_errata.pdf

[10] In this datasets, Japanese texts were translated from English broadcast conversations. Despite the multifold domain mismatch (i.e., spoken and translationese), to our knowledge, this was the best dataset available for our purpose.

Methods	Web			
	all	*intra*	*inter*	*exophora*
Ueda et al. (2020)[9]	70.3	-	-	-
+MLM	71.0 ±0.716	63.9 ±1.27	65.1 ±1.14	75.8 ±0.764
+MT	70.5 ±0.410	64.0 ±0.868	63.8 ±0.536	75.4 ±0.565
+MT w/ MLM	**71.9** ±0.416	**65.4** ±0.697	**65.2** ±1.07	**76.7** ±0.468
+MT (Two-stage)	71.4 ±0.511	65.3 ±0.830	64.6 ±0.479	76.1 ±0.633
+MT w/ MLM (Two-stage)	71.7 ±0.393	65.0 ±0.641	64.4 ±0.844	**76.7** ±0.478

Table 3: Breakdown of the F1 scores with standard deviations on the **Web** test set. Boldfaced scores indicate the best results in the corresponding categories. One-stage optimization with the MLM objective performed the best on all categories.

Methods	News			
	all	*intra*	*inter*	*exophora*
Ueda et al. (2020)[9]	56.7	-	-	-
+MLM	57.1 ±0.359	62.7 ±0.723	50.2 ±0.880	55.6 ±1.30
+MT	57.7 ±0.442	63.8 ±0.652	49.8 ±0.811	**57.1** ±0.910
+MT w/ MLM	**58.3** ±0.383	**65.0** ±0.544	**50.6** ±0.667	56.3 ±1.07
+MT (Two-stage)	57.3 ±0.466	63.2 ±0.723	50.1 ±0.761	55.7 ±0.700
+MT w/ MLM (Two-stage)	57.7 ±0.549	63.8 ±0.597	50.2 ±0.628	56.2 ±1.37

Table 4: Breakdown of the F1 scores with standard deviations on the **News** test set. Boldfaced scores indicate the best results in the corresponding categories. One-stage optimization with the MLM objective performed the best on all categories but exophora.

	Web	News
intra-sentential anaphora	0.758	0.763
inter-sentential anaphora	0.871	0.879

Table 5: Pearson's correlation coefficient between ZPT accuracies and ZAR F1 scores.

scores were averaged over 3 different seeds.

Through the course of intermediate training, we observed almost steady increase in ZPT accuracies and ZAR F1 scores until around the 30th epoch (the four figures in Appendix D). Table 5 shows the strong positive correlations between the two performance measures, especially the very strong correlation for *inter-sentential* anaphora. These results were in line with our speculation that the performance gains in ZAR stemmed from the model's increased ability to translate zero pronouns.

4.5 Why Is MLM so Effective?

The MLM objective during intermediate training on MT is shown to be very effective, but why? Pruksachatkun et al. (2020) conjecture that it would mitigate catastrophic forgetting (CF), but this is not the sole explanation. In fact, Konno et al. (2020) see token masking as a way to augment data.

Methods	Web		News	
	F1	△	F1	△
+MT	70.5	-	57.7	-
+MT w/ masking	71.1	**0.6**	57.8	0.1
+MT w/ MLM	71.9	**1.4**	58.3	0.6

Table 6: Ablation study focusing on MLM. All models were trained with one-stage optimization. **w/ masking** indicates token masking without the corresponding loss function.

To dig into this question, we conducted an ablation study by introducing a model with token masking but without the corresponding loss function (denoted as **+MT w/ masking**). We assume that this model was largely deprived of the power to mitigate CF while token masking still acted as a data augmenter.

Table 6 shows the results. Not surprisingly, **+MT w/ masking** was beaten by **+MT w/ MLM** with large margins. However, it did outperform **+MT**, and the gain was particularly large for the Web. The fact that the contribution of the loss function was larger than that of token masking indicates that the improvements were mainly attributed to CF mitigation, but the contribution of token masking alone should not be overlooked.

4.6 Case Studies

To gain further insights, we compared ZAR results with English translations automatically generated by the corresponding MT model. Figure 4 gives two examples. It is no great surprise that the translation quality was not satisfactory because we did not fully optimize the model for it.

In the exmple of Figure 4(a), MT seems to have helped ZAR. The omitted nominative argument of "あり" (is) was correctly translated as "**the school**", and the model successfully identified its antecedent "学校" (school) while the baseline failed.

Figure 4(b) illustrates a limitation of the proposed approach. The omitted nominative argument of the predicate, "で" (be), points to "定吉" (Sadakichi, the father of Jutaro). Although the model correctly translated the zero pronoun as "**He**", it failed in ZAR. This is probably because not only "定吉 (Sadakichi)" but also "龍馬" (Ryoma) and "重太郎" (Jutaro) can be referred to as "He". When disambiguation is not required to generate an overt pronoun, MT is not very helpful.

4.7 Note on Other Pretrained Models

Due to space limitation, we have limited our focus to BERT, but for the sake of future practitioners, we would like to briefly note that we extensively tested BART (Lewis et al., 2020) and its variants before switching to BERT. Unlike BERT, BART is an encoder-decoder model pretrained on a monolingual corpus (original) or a non-parallel multilingual corpus (mBART) (Liu et al., 2020). Because MT requires the encoder-decoder architecture, maintaining the model architecture between pretraining and intermediate training looked promising to us.

We specifically tested (1) the officially distributed mBART model, (2) a BART model we pretrained on Japanese Wikipedia, and (3) an mBART model we pretrained on Japanese and English texts. During fine-tuning, we added the ZAR argument selection layer on top of either the encoder or the decoder.

Unfortunately, gains from MT intermediate training were marginal for these models. A more serious problem was that they came close to but rarely outperformed the strong BERT baseline. We gave up identifying the cause of poorer performance because it was extremely hard to apply comparable experimental conditions to large pretrained models.

5 Conclusion

In this paper, we proposed to exploit parallel texts for Japanese zero anaphora resolution (ZAR) by inserting machine translation (MT) as an intermediate task between masked language model (MLM) pretraining and fine-tuning on ZAR. Although previous studies reported negative results on the use of MT as an intermediate task, we demonstrated that it did work for Japanese ZAR. Our analysis suggests that the intermediate training on MT simultaneously improved the model's ability to translate Japanese zero pronouns and the ZAR performance.

We bridged the gap between BERT-based ZAR and the encoder-decoder architecture for MT by initializing the encoder part of the MT model with a pretrained BERT. Previous studies focusing on MT reported mixed results on this approach, but again, we demonstrated its considerable positive impact on ZAR. We found that incorporating the MLM objective into the intermediate training was particularly effective. Our experimental results were consistent with the speculation that MLM mitigated catastrophic forgetting during intermediate training.

With neural transfer learning, we successfully revived the old idea that Japanese ZAR can benefit from parallel texts (Nakaiwa, 1999). Thanks to the astonishing flexibility of neural networks, we would probably be able to connect ZAR to other tasks through transfer learning.

Acknowledgments

We thank the Yomiuri Shimbun for providing Japanese-English parallel texts. We are grateful for Nobuhiro Ueda' help in setting up the baseline model. We thank the anonymous reviewers for their insightful comments.

References

Stephane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. 2019. On the use of BERT for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 108–117, Hong Kong. Association for Computational Linguistics.

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*

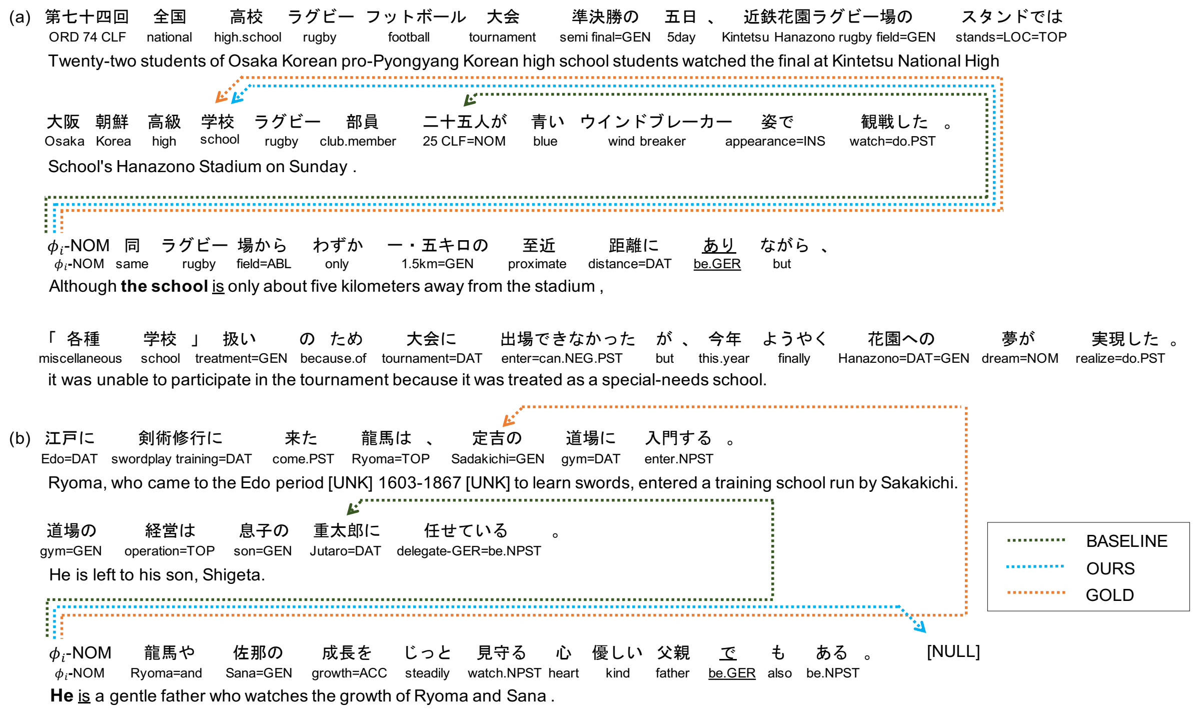

(a) 第七十四回 全国 高校 ラグビー フットボール 大会 準決勝の 五日 、 近鉄花園ラグビー場の スタンドでは
ORD 74 CLF national high.school rugby football tournament semi final=GEN 5day Kintetsu Hanazono rugby field=GEN stands=LOC=TOP
Twenty-two students of Osaka Korean pro-Pyongyang Korean high school students watched the final at Kintetsu National High
大阪 朝鮮 高級 学校 ラグビー 部員 二十五人が 青い ウインドブレーカー 姿で 観戦した 。
Osaka Korea high school rugby club.member 25 CLF=NOM blue wind breaker appearance=INS watch=do.PST
School's Hanazono Stadium on Sunday .
ϕ_i-NOM 同 ラグビー 場から わずか 一・五キロの 至近 距離に あり ながら 、
ϕ_i-NOM same rugby field=ABL only 1.5km=GEN proximate distance=DAT be.GER but
Although the school is only about five kilometers away from the stadium ,
「 各種 学校 」 扱い の ため 大会に 出場できなかった が 、 今年 ようやく 花園への 夢が 実現した 。
miscellaneous school treatment=GEN because.of tournament=DAT enter=can.NEG.PST but this.year finally Hanazono=DAT=GEN dream=NOM realize=do.PST
it was unable to participate in the tournament because it was treated as a special-needs school.
(b) 江戸に 剣術修行に 来た 龍馬は 、 定吉の 道場に 入門する 。
Edo=DAT swordplay training=DAT come.PST Ryoma=TOP Sadakichi=GEN gym=DAT enter.NPST
Ryoma, who came to the Edo period [UNK] 1603-1867 [UNK] to learn swords, entered a training school run by Sakakichi.
道場の 経営は 息子の 重太郎に 任せている 。
gym=GEN operation=TOP son=GEN Jutaro=DAT delegate-GER=be.NPST
He is left to his son, Shigeta.
ϕ_i-NOM 龍馬や 佐那の 成長を じっと 見守る 心 優しい 父親 で も ある 。 [NULL]
ϕ_i-NOM Ryoma=and Sana=GEN growth=ACC steadily watch.NPST heart kind father be.GER also be.NPST
He is a gentle father who watches the growth of Ryoma and Sana .
BASELINE
OURS
GOLD

Figure 4: Two examples of ZAR and MT. Green, blue, and orange dotted lines represent the output of the baseline model, that of ours, and the gold standard, respectively. English sentences are generated by the corresponding MT (encoder-decoder) model. (a) The example in which MT apparently helped ZAR. The nominative zero pronoun of "あり" (is) was correctly translated as "**the school**". The model also succeeded in identifying its antecedent "学校" (school). (b) The example in which MT was not helpful. The model successfully translated the nominative zero pronoun of the underlined predicate, "で" (be), as "**He**". It misidentified its antecedent, however.

Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tomomasa Furukawa, Toshiaki Nakazawa, Tomohide Shibata, Daisuke Kawahara, and Sadao Kuroahshi. 2017. Automatic construction of a pseudo-annotatedzero anaphora corpus using a bilingual corpus. In *Proceedings of the Twenty-third Annual Meeting of the Association for Natural Language Processing*. (in Japanese).

Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, and Enhong Chen. 2020. Incorporating BERT into parallel sequence decoding with adapters. In *Advances in Neural Information Processing Systems*, volume 33, pages 10843–10854. Curran Associates, Inc.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2012. Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 535–544, Bali, Indonesia. Faculty of Computer Science, Universitas Indonesia.

Ryu Iida, Mamoru Komachi, Naoya Inoue, Kentaro Inui, and Yuji Matsumoto. 2017. NAIST text corpus: Annotating predicate- argument and coreference relations in Japanese. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 1177–1196. Springer, Dordrecht.

Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, Canasai Kruengkrai, and Julien Kloetzer. 2016. Intra-sentential subject zero anaphora resolution using multi-column convolutional neural network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1244–1254, Austin, Texas. Association for Computational Linguistics.

Kenji Imamura and Eiichiro Sumita. 2019. Recycling a pre-trained BERT encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31, Hong Kong. Association for Computational Linguistics.

Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. 2002. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Youngtae Kim, Dongyul Ra, and Soojong Lim. 2021. Zero-anaphora resolution in Korean based on deep language representation model: BERT. *ETRI Journal*, 43(2):299–312.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Ryuto Konno, Shun Kiyono, Yuichiroh Matsubayashi, Hiroki Ouchi, and Kentaro Inui. 2021. Pseudo zero pronoun resolution improves zero anaphora resolution. *arXiv:2104.07425*.

Ryuto Konno, Yuichiroh Matsubayashi, Shun Kiyono, Hiroki Ouchi, Ryo Takahashi, and Kentaro Inui. 2020. An empirical study of contextual data augmentation for Japanese zero anaphora resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4956–4968, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Shuhei Kurita, Daisuke Kawahara, and Sadao Kurohashi. 2018. Neural adversarial training for semi-supervised Japanese predicate-argument structure analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 474–484, Melbourne, Australia. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Ting Liu, Yiming Cui, Qingyu Yin, Wei-Nan Zhang, Shijin Wang, and Guoping Hu. 2017. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 102–111, Vancouver, Canada. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kojiro Nabeshima and Michael N. Brooks. 2020. *Techniques of English-Japanese translation*. Kurosio Publishers. (in Japanese).

Masaaki Nagata and Makoto Morishita. 2020. A test set for discourse translation from Japanese to English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3704–3709, Marseille, France. European Language Resources Association.

Hiromi Nakaiwa. 1999. Automatic extraction of rules for anaphora resolution of Japanese zero pronouns in Japanese-English machine translation from aligned sentence pairs. *Machine Translation*, 14(14):247–279.

Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2017. Neural modeling of multi-predicate interactions for Japanese predicate argument structure analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1591–1600, Vancouver, Canada. Association for Computational Linguistics.

Arum Park, Seunghee Lim, and Munpyo Hong. 2015. Zero object resolution in Korean. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 439–448, Shanghai, China.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. *arXiv:1811.01088*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

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 (Coling 2008)*, pages 769–776, Manchester, UK. Coling 2008 Organizing Committee.

Ryohei Sasano and Sadao Kurohashi. 2011. A discriminative approach to Japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 758–766, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. 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 (Volume 1: Long Papers)*, pages 579–589, Melbourne, Australia. 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*, pages 3630–3634, Marseille, France. European Language Resources Association.

Linfeng Song, Kun Xu, Yue Zhang, Jianshu Chen, and Dong Yu. 2020. ZPR2: Joint zero pronoun recovery and resolution using multi-task learning and BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5429–5434, Online. Association for Computational Linguistics.

Nobuhiro Ueda, Daisuke Kawahara, and Sadao Kurohashi. 2020. BERT-based cohesion analysis of Japanese texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1323–1333, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019a. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Xing Wang, and Shuming Shi. 2019b. 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 (EMNLP-IJCNLP)*, pages 921–930, Hong Kong, China. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2018. Learning to jointly translate and predict dropped pronouns with a shared reconstruction mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2997–3002, Brussels, Belgium. Association for Computational Linguistics.

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*, pages 983–993, San Diego, California. Association for Computational Linguistics.

Souta Yamashiro, Hitoshi Nishikawa, and Takenobu Tokunaga. 2018. Neural Japanese zero anaphora resolution using smoothed large-scale case frames with word embedding. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Qingyu Yin, Yu Zhang, Weinan Zhang, and Ting Liu. 2017. Chinese zero pronoun resolution with deep memory network. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1309–1318, Copenhagen, Denmark. Association for Computational Linguistics.

Jia-Rui Zhang, Hongzheng Li, Shumin Shi, Heyan Huang, Yue Hu, and Xiangpeng Wei. 2020. Dynamic attention aggregation with bert for neural machine translation. In *IJCNN*, pages 1–8. IEEE.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating BERT into neural machine translation. In *International Conference on Learning Representations*.

A Hyper-parameters for MT

Options	Values
Optimizer	Adam
Adam params	β_1=0.9, $\beta_2 = 0.98$
Optimizer eps	1×10^{-6}
Weight decay	0.01
Epochs	50
First-stage epochs*	15
Batch size	Approx. 500
Learning rate	3.0×10^{-4}
Warm-up	5 epochs
Loss function	Lable-smoothed cross entropy
Label smoothing	0.1
Dropout (BERT & Dec.)	0.1
LR Scheduler	polynomial decay

Table 7: Hyper-parameters for MT. *: For two-stage optimization.

B Hyper-parameters for ZAR

Options	Values
Optimizer	AdamW
Optimizer eps	1×10^{-8}
Weight decay	0.01
Epochs	4
Batch size	8
Learning rate	5.0×10^{-5}
Warmup proportion	0.1
Loss function	Cross entropy loss
Dropout (BERT layer)	0.1
Dropout (output layer)	0.0
LR Scheduler	linear_schedule- _with_warmup [11]

Table 8: Hyper-parameters for ZAR.

Although we followed Ueda et al. (2020) with respect to hyper-parameter settings, there was one exception. Verbal predicate analysis is conventionally divided into three types: *overt*, *covert*, and *zero*. While Ueda et al. (2020) excluded the easiest *overt* type from training, we targeted all the three types because we found slight performance improvements. The *overt* type covers situations where the predicate and the argument have a dependency relation and their relation is marked explicitly with a case marker.

C Results on Validation Sets

Tables 9 and 10 show the performance on the validation sets.

D Relationship between Zero Anaphora Resolution and Zero Pronoun Translatoin

Figure 5 shows the relationship between zero anaphora resolution (ZAR) and zero pronoun translation (ZPT) in the course on intermediate training on MT. We observed almost steady increase in ZPT accuracies and ZAR F1 scores until around the 30th epoch.

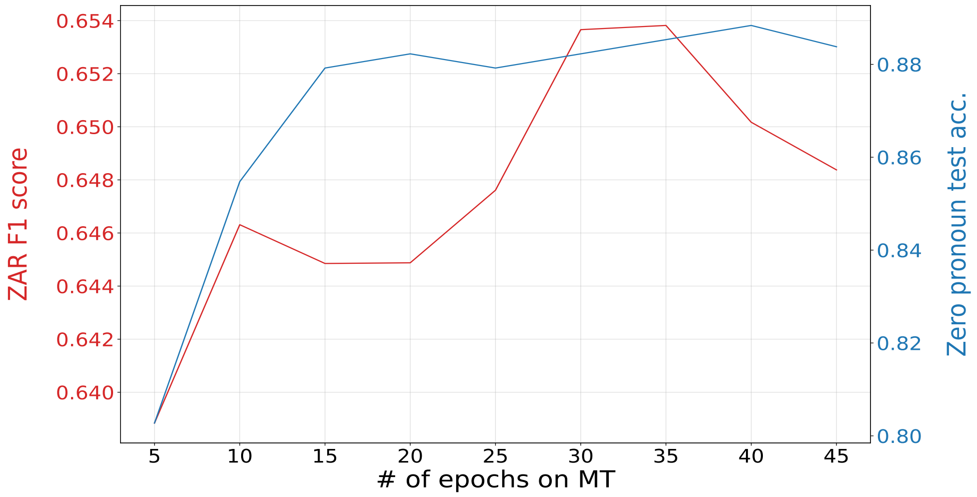
[11] https://github.com/huggingface/transformers/blob/v2.10.0/src/transformers/optimization.py#L47

Methods	Web			
	all	*intra*	*inter*	*exophora*
+MLM	62.9 ±0.812	56.8 ±1.03	53.5 ±0.827	68.5 ±0.962
+MT	62.9 ±0.668	56.6 ±0.830	51.4 ±0.875	68.9 ±0.825
+MT w/ MLM	**64.1** ±0.475	**58.4** ±0.883	52.9 ±2.01	69.8 ±0.993
+MT (Two-stage)	63.6 ±0.437	57.5 ±1.36	52.7 ±2.15	69.4 ±0.514
+MT w/ MLM (Two-stage)	63.9 ±0.488	57.4 ±1.07	**53.1** ±1.56	**69.9** ±0.831

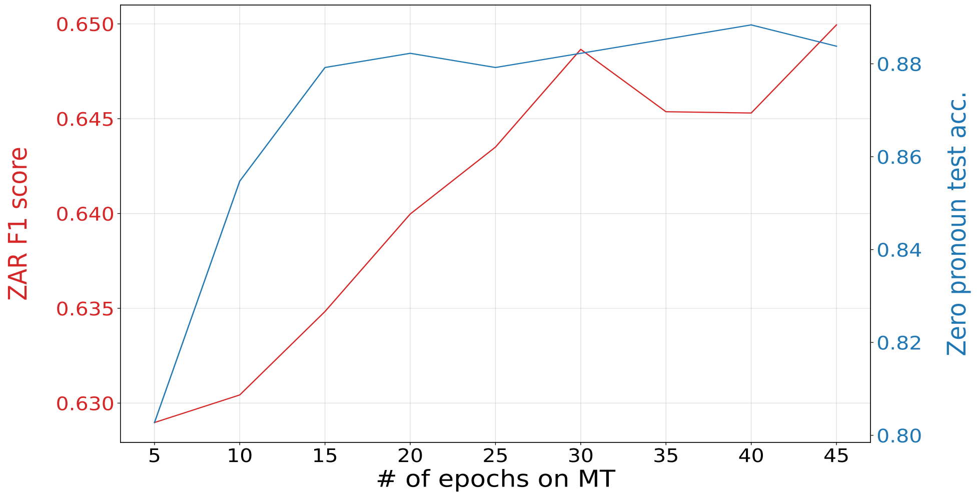
Table 9: Breakdown of the F1 scores with standard deviations on the **Web** validation set. Boldfaced scores indicate the best results in the corresponding categories.

Methods	News			
	all	*intra*	*inter*	*exophora*
+MLM	57.8 ±0.586	64.1 ±0.80	48.1 ±0.965	56.3 ±1.81
+MT	57.0 ±0.710	62.9 ±0.819	46.0 ±1.23	**58.3** ±1.80
+MT w/ MLM	**58.7** ±0.438	**65.5** ±1.08	48.3 ±0.843	57.6 ±1.48
+MT (Two-stage)	58.4 ±0.579	64.1 ±0.714	**48.8** ±1.57	57.9 ±1.20
+MT w/ MLM (Two-stage)	58.6 ±0.381	64.9 ±0.543	48.7 ±0.767	58.0 ±1.16

Table 10: Breakdown of the F1 scores with standard deviations on the **News** validation set. Boldfaced scores indicate the best results in the corresponding categories.

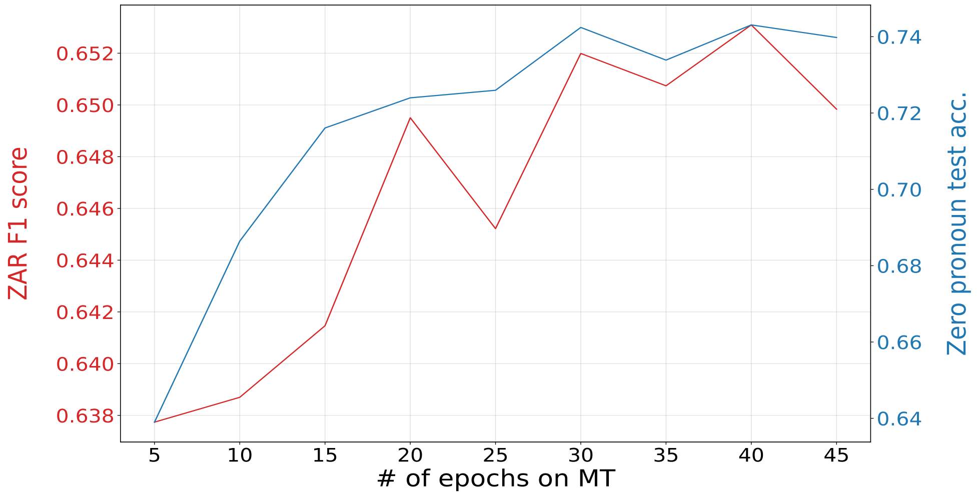


(a) Relationship between ZAR and ZPT for *intra-sentential* on the **Web** test set.

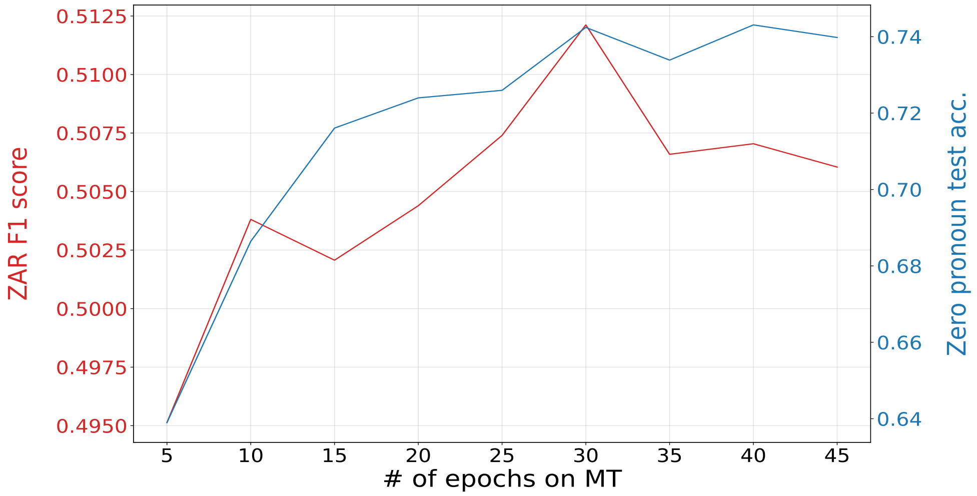


(b) Relationship between ZAR and ZPT for *intra-sentential* on the **News** test set.

Figure 5: Relationships between ZAR and ZPT.



(c) Relationship between ZAR and ZPT for *inter-sentential* on the **Web** test set.



(d) Relationship between ZAR and ZPT for *inter-sentential* on the **News** test set.

Figure 5: Relationships between ZAR and ZPT.