GuoFeng: A Benchmark for Zero Pronoun Recovery and Translation

Longyue Wang[1*] **Mingzhou Xu**[2*] **Derek F. Wong**[2] **Hongye Liu**[1]
Linfeng Song[1] **Lidia S. Chao**[2] **Shuming Shi**[1] **Zhaopeng Tu**[1]
[1]Tencent AI Lab
[2]University of Macau
{vinnylywang, hongyeliu, lfsong, shumingshi, zptu}@tencent.com
nlp2ct.mzxu@gmail.com, {derekfw, lidiasc}@um.edu.com

Abstract

The phenomenon of zero pronoun (ZP) has attracted increasing interest in the machine translation (MT) community due to its importance and difficulty. However, previous studies generally evaluate the quality of translating ZPs with BLEU scores on MT testsets, which are not expressive or sensitive enough for accurate assessment. To bridge the data and evaluation gaps, we propose a *benchmark testset* for target evaluation on Chinese-English ZP translation. The human-annotated testset covers five challenging genres, which reveal different characteristics of ZPs for comprehensive evaluation. We systematically revisit eight advanced models on ZP translation and identify current challenges for future exploration. We release data, code, models and annotation guidelines, which we hope can significantly promote research in this field.[1]

1 Introduction

Zero pronoun (ZP) is a discourse phenomenon that appears frequently in pronoun-dropping (pro-drop) languages such as Chinese and Japanese. Specifically, pronouns are often omitted when they can be pragmatically or grammatically inferred from intra- and inter-sentential contexts (Li and Thompson, 1979). Since recovery of such ZPs generally fails, this poses difficulties for several generation tasks, including dialogue modelling (Su et al., 2019), question answering (Tan et al., 2021) as well as machine translation (MT) (Wang et al., 2018a).

Although neural machine translation (NMT) (Bahdanau et al., 2015; Vaswani et al., 2017a; Gehring et al., 2017) has achieved great progress in recent years, translating ZPs from a pro-drop to a non-pro-drop language is a significant and challenging task (Wang et al., 2018a). As shown in Table 1, an advanced NMT system still fails to recall the ZP information, which leads to severe problems in translation outputs: 1) *incompleteness* where the first case grammatically lacks the subject "She", and the associated verb should be "teaches"; and 2) *incorrectness* where the second case is semantically incorrect (i.e., it should be "let her buy it" instead of "let you buy") due to unresolved zero anaphora of "她" and "它".

Inp.	黄娟，女，副教授 。 (她) 主要 教授 《商务英语》 课程 。
Out.	Huang Juan, female, associate professor. Mainly teach the course Business English.
Inp.	A: 菲比 很 想 买 台 电视 。 B: 乔伊 不 让 (她) 买 (它) ？
Out.	A: Phoebe really wants to buy a TV. B: Joey won't let you buy?

Table 1: Examples of ZP translation from two different genres. Words in brackets are ZPs that are invisible in decoding and underlined words are corresponding antecedents. "Inp." and "Out." represent the Chinese input and output of *Google Translate*, respectively. As seen, the translations are either incomplete or incorrect.

In recent years, there has been a growing interest in zero pronoun translation (ZPT), which aims to directly recover ZPs in the translation (Wang et al., 2016, 2018a; Yu et al., 2020; Ri et al., 2021). Most studies only report performance on MT testsets in terms of BLEU scores, which is not expressive or sensitive enough to capture the translation quality on ZPs. To bridge this gap, we propose a Chinese-English **benchmark testset** (mZPRT) specially targeting ZPT, which has the following appealing characteristics:

- **Human Annotation**: The current ZPT dataset is automatically annotated using alignment information that is not accurate enough (i.e., TVsub dataset (Wang et al., 2018a)). Ours is built by professional annotators and translators, which can

*Mingzhou Xu and Longyue Wang contributed equally to this work. Work was done when Mingzhou Xu and Hongye Liu were interning at Tencent AI Lab.

[1]https://github.com/longyuewangdcu/mZPRT.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11266 - 11278
December 7-11, 2022

Dataset	HA	MD	ZPR	ZPT
OntoNotes (Pradhan et al., 2012)	✓	✓	✗	✗
BaiduKnows (Zhang et al., 2019)	✓	✗	✓	✗
TVsub (Wang et al., 2018a)	✗	✗	✗	✓
This Work	✓	✓	✓	✓

Table 2: Comparison of the proposed dataset to existing ones regarding ZP phenomenon. "HA" and "MD" represent whether a corpus is human-annotated and multi-domain, respectively. "ZPR" and "ZPT" indicate whether a dataset is widely used for evaluating specific tasks. The symbol ✓ or ✗ means "Yes" or "No".

provide more accurate and relevant indications of model performance.

- **Multiple Domains**: The frequencies and types of ZPs vary in different domains (Yang et al., 2015) but previous studies only consider a single domain. Our testset covers five challenging domains, which can comprehensively reveal different characteristics in respective domains.
- **Annotations for Both ZP Recovery and Translation**: There is no dataset consisting of both source ZPs and target translations. We annotate both of them in our dataset, which enables explicit investigation of the effects of recovering ZPs on the performance of ZP translation.

The dataset is significantly different from existing ones, as listed in Table 2. Experimental results on the proposed benchmark show significant differences in model behavior and quality across domains, emphasizing the need for more comprehensive evaluation as a standard procedure. Besides, a good external system of ZP recovery can benefit ZP translation, but there is still a large space for further improvement. Through revisiting recent advanced models from the perspective of ZP translation, we obtain some interesting findings:

- Scaling NMT models can achieve great performance gains in terms of BLEU score, however, the translation accuracy regarding ZPs still can not be guaranteed.
- Although BLEU scores do not vary significantly, we prove that document-level NMT models are helpful to ZP translation.

2 Preliminaries

2.1 Zero Pronoun in Machine Translation

The anaphora phenomenon is the meaning of an element depends on the antecedent element in context, which can be considered one of the most challenging problems in natural language processing, especially for ZP (Peral and Ferrández, 2003). Recent years have seen a surge of interest in zero pronoun recovery (ZPR) and ZPT tasks, which respectively resolve ZPs in the source and target language (e.g. Chinese⇒English).

Zero Pronoun Recovery Given a source sentence, ZPR aims to insert dropped pronouns in proper positions without changing the original meanings (Yang and Xue, 2010; Yang et al., 2015, 2019). It is different from the task of ZP resolution, which identifies the antecedent of a referential pronoun (Mitkov, 2014). However, more than 50% ZPs are non-anaphoric (Rao et al., 2015), which is not directly helpful to MT task compared with ZPR systems. Previous studies regarded ZPR as a classification or sequence labeling problem, and advanced models can only achieve 40∼60% F1 scores on closed datasets (Zhang et al., 2019), indicating the difficulty of understanding zero anaphora.

Zero Pronoun Translation When pronouns are omitted in a source sentence, ZPT aims to generate ZPs in its target translation. Generally, prior works fall into three categories: 1) *pipeline*, where input sentences are labeled with ZPs using an external ZPR system and then fed into a standard MT model (Chung and Gildea, 2010; Wang et al., 2016, 2017b); 2) *implicit*, where ZP phenomenon is implicitly resolved by modelling document-level contexts (Wang et al., 2017a; Yu et al., 2020; Ri et al., 2021); 3) *end-to-end*, where ZP prediction and translation are jointly learned in an end-to-end manner (Wang et al., 2018a,b, 2019; Tan et al., 2021). Although these methods have achieved performance improvement to some extent, they still face a few major weaknesses.

2.2 Discussion and Motivation

Lack of Comprehensive Benchmark Testsets Two technological advances in the field of ZPR and ZPT have seen vast progress over the last decades, but they have been developed very much in isolation. For instance, the ZPR systems are mainly trained and evaluated on the human-labeled BaiduKnows Corpus while the ZPT task has experimented on the auto-annotated TVsub. This critical data gap severely limits the investigation of the effects of ZPR on the performance of ZPT. Furthermore, the frequencies and types of ZPs vary in different genres (Yang et al., 2015). The most

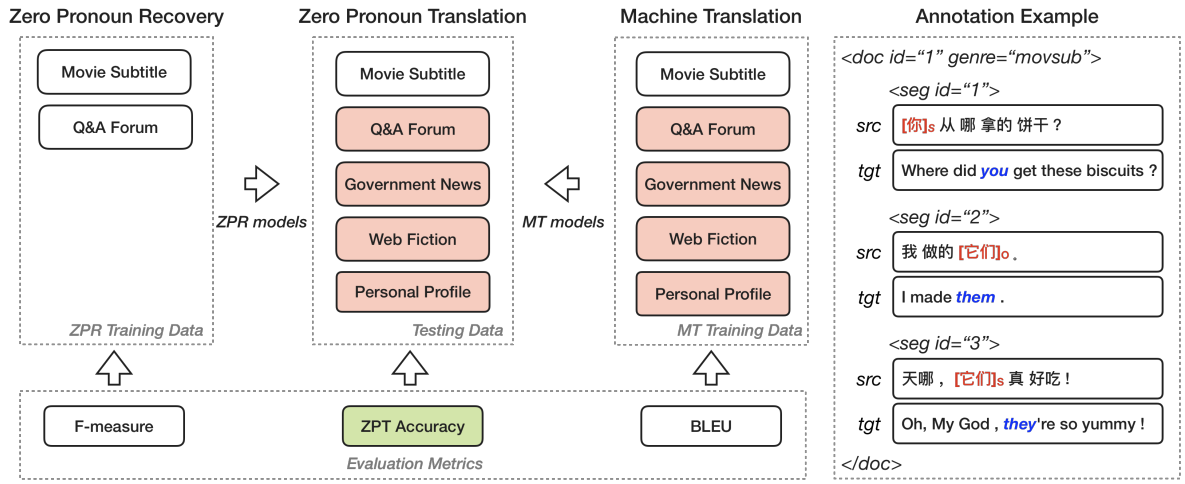

Figure 1: An overview of the proposed multi-domain dataset and evaluation metric. They can be used for evaluating MT, ZPR and ZPT tasks, which were often benchmarked in isolation. The white boxes indicate existing resources, while colored ones are newly proposed or collected in this work. We also show an example with annotations.

frequent ZP in newswire text is the third person singular 它 ("it") (Baran et al., 2012), while that in SMS dialogues is the first person 我 ("I") and 我们 ("we") (Rao et al., 2015). This may lead to differences in model behavior and quality across domains. Thus, experimental results are neither reliable nor comprehensive when evaluated on a domain-specific dataset. Motivated by this, we propose a multi-domain benchmark dataset for evaluating ZPR and ZPT tasks.

Lack of Fine-Grained Evaluation Previous works usually evaluate ZPT models using the BLEU metric (Wang et al., 2016, 2018a; Yu et al., 2020; Ri et al., 2021), however, we empirically found that general-purpose metrics cannot characterize the performance of ZP translation. As shown in Table 1, the missed or incorrect pronouns may not affect BLEU scores but severely harm true performances. To bridge this gap, we also measure model performance using a pronoun-specific evaluation metric (Guillou and Hardmeier, 2018). To this end, we systematically compared existing ZPT methods, and highlighted advances in ZPT quality that go beyond BLEU improvement.

3 mZPRT: New Benchmark Dataset

In general, the process of manual labeling totally spends five annotators and two translators two months, which costs US $10,000 dollars.

3.1 Data Source

We determined five domains of texts (i.e. movie subtitle, Q&A forum, government news, web fiction, and personal profile) that contain a proportion of ZPs. Accordingly, we construct these subsets in three ways: 1) *ZP Labeling*, where expert annotators annotate the source side of a parallel dataset with fine-grained ZP labels; 2) *Translating*, where professional translators extend a monolingual ZP-labeled dataset into a parallel one by translating Chinese sentences into English; 3) *Both*, where we build them from scratch (e.g. labeling, translating).

- *Movie Subtitle.* This is collected from validation and testsets of the TVsub corpus,[2] which has been auto-annotated with coarse-grained ZP labels and translations. Then, we checked and re-annotated ZPs with fine-grained labels.
- *Q&A Forum.* The source-side sentences are collected from the testset of the Baidu Knows corpus,[3] which has been annotated with coarse-grained ZP labels with boundary tags. We extend them into parallel data and re-annotated ZPs with fine-grained labels.
- *Government News.* We crawled relevant texts from the bilingual news website[4] and then manually aligned sentences. The source-side data are labeled with ZPs, and each article is regarded as one document.
- *Web Fiction.* We crawled 24 chapters from 5 genres of books in Chinese[5] and English[6] Webnovel websites. We manually aligned Chinese-English

[2] github.com/longyuewangdcu/tvsub
[3] zhidao.baidu.com
[4] language.chinadaily.com.cn
[5] www.qidian.com
[6] www.webnovel.com

ID	Subset	Agree.	Size (#)			ZP Type Dist. (%)			ZP Sent. Freq. (%)	
			|D|	|S|	|W|	Sub.	Obj.	Pos.	≥ 1	≥ 2
1	Movie Subtitle	0.90	25	2,204	13K/18K	65.7	15.4	16.9	36.3	6.7
2	Q&A Forum	0.86	182	1,171	16K/22K	90.9	4.2	4.8	44.0	12.7
3	Government News	0.95	8	1,587	36K/51K	82.1	2.0	15.9	56.5	13.2
4	Web Fiction	0.92	24	1,658	33K/39K	66.3	8.0	24.1	39.2	12.2
5	Personal Profile	0.94	218	1,473	47K/59K	93.1	0.1	6.8	49.3	7.8
Total / Average		0.91	457	8,093	146K/190K	79.6	5.9	13.7	44.3	10.2

Table 3: Statistics of our dataset, covering five subsets in the respective domains. First, we compute the inter-annotator agreement via pairwise kappa coefficient (Agree.). Second, we count the number of documents (|D|), sentences (|S|) and words (|W|). Third, we categorize ZPs into five forms (detailed in Table 10) and report distribution of three main types: subject, object, and possessive adjective (ZP Type Dist.). To show the frequency of the ZP phenomenon, we calculate the percentage of sentences that contain more than one / two ZPs (ZP Sent. Freq.)

.

sentences to build a parallel version. We labeled ZPs and then tagged document boundaries according to chapter information.

- *Personal Profile.* The source-side texts are collected from 218 homepages of academic staff, covering 50 academic majors in QS2021 top-10 universities in China. We translate Chinese sentences into English and labeled ZPs on the source side. The texts from the same homepage can be regarded as one document.

3.2 Human Annotation

Taking an annotation example in Figure 1 for instance, the annotation guidelines are as follows:

1. In a Chinese sentence, ZP can be ascertained by checking its syntactic structure. To avoid annotation ambiguity, we provide English translation for reference. Since English is a non-pro-drop language, Chinese ZP can be detected according to its English equivalent (e.g. *you* $\Rightarrow$ [你]).
2. The anaphoric ZPs can be recovered by considering their antecedents in the context (e.g. 饼干 $\Rightarrow$ [它们]). To improve the annotation accuracy, the ZPs should be double-checked using English equivalents (e.g. *them* $\Rightarrow$ biscuits $\Rightarrow$ [它们] instead of [他们] or [她们]).
3. The non-anaphoric ZPs can be recovered by inferring from salient entities in the environment (e.g. [你] $\Rightarrow$ *The Hearer*). If possible, this should be double-checked using English equivalents (e.g. *you* $\Rightarrow$ [你] instead of [我].)
4. We additionally annotate ZPs with fine-grained labels (e.g. $[你]_s$) according to their forms in the sentence, including subject, object, possessive adjective and reflexive (i.e. $[\cdot]_s$, $[\cdot]_o$, $[\cdot]_p$, $[\cdot]_r$). This information is helpful to fine-grained analysis (in Section 3.3) and automatic evaluation (in Section 4.1).
5. Some subsets only contain Chinese sentences, which need to be translated into English. Translators could consider document-level contexts to generate discourse-aware translations (i.e. sentences #1 and #2 are given when translating sentence #3, thus human translators can restore "they" in the target language).
6. The document boundaries are also tagged according to content clues such as topic, storyline and scene (e.g. <doc>. . . </doc>).

The "Personal Profile" subset was collected from academic homepages, which are public on the University websites. It does not contain any sensitive information such as ID number, phone number, home address, email, etc. About the personal profile (e.g. person name, school name and publication), we have replaced them according to the following process:

1. Person names are replaced with Person-A, . . . ;
2. Organization names are replaced with University-of-A, College-of-B, School-of-C, Faculty-of-D, . . . ;
3. Numbers (i.e. year of work experience and number of publication) are randomly shuffled.

About ZP annotation, each individual annotates all sentences in one subset and 50 overlapping sentences sampled from the other four subsets. We followed Mitani et al. (2017) to measure the inter-annotator agreement by calculating the average pairwise cohen's kappa score. As shown in Table 3, our dataset reaches 0.86%$\sim$0.95% kappa

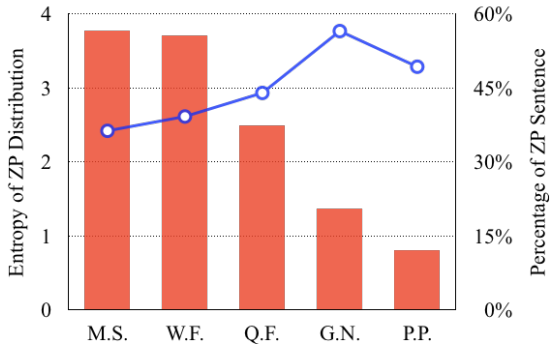

Figure 2: Analysis of our dataset across five domains. Red histogram demonstrates label density by calculating the entropy of ZP distribution. Blue polyline indicates the frequency of the phenomenon by computing the percentage of sentences containing ZPs.

scores (i.e. 0.91% on average), demonstrating that the annotators work efficiently and consistently under this guideline. We attribute this to the fact that English translation is more helpful to disambiguate Chinese ZPs than only considering syntactic and semantic information in a single language. This is potentially useful for annotating ZP datasets in other languages.

3.3 Data Statistics and Analysis

As illustrated in Table 3, the proposed dataset contains 457 documents with over 8,000 Chinese⇒English parallel sentences. According to our fine-grained labels, we also report ZP distribution in terms of three main types: subject, object, and possessive adjective. As seen, ZPs in two domains (Q&A forum and personal profile) are subject types, covering more than 90% of all ZPs. On the contrary, web fiction contains more possessive adjective ZPs (24.1%) while ZP types in movie subtitles distribute more dispersedly. For the average value across domains, subject ZPs still occupy a large proportion while object types only occupy a small proportion (78% vs. 5%).

Figure 2 further analyzes different characteristics of our testset across domains. On the one hand, we demonstrate label density by calculating the entropy of fine-grained ZP distribution. By assuming that high density indicates the difficulty of modeling ZPs, movie subtitle and web fiction are more challenging than other domains. On the other hand, we report the frequency of the phenomenon by computing the percentage of sentences containing ZPs. As seen, government news and personal profile contain many more ZPs than other domains.

Domain	ZPR	$MT_{SEN.}$	$MT_{DOC.}$	$MT_{REC.}$
Mov. Subtitle	2.2M	12.8M	12.8M	2.2M
Q&A Forum	5K	50K	50K	2.2M
Web Fiction	2.3M	1.5M	1.5M	2.2M
Gov. News	2.3M	42.8M	42.8M	2.2M
Per. Profile	2.3M	42.8M	42.8M	2.2M

Table 4: Domain and size of training data for building ZPR, sentence-level, document-level and reconstruction-based NMT models. The cell colors denote in-domain, out-of-domain and mix-domain. The size is counted in number of sentence pairs.

In summary, 1) this emphasizes the necessity of evaluating ZPT across multiple domains; 2) it reveals the label imbalance and sparsity problems in ZP-aware tasks.

4 Experiment

We systematically evaluate existing models on the proposed mZPRT benchmark to 1) investigate effects of ZPR on ZPT; 2) build a benchmark of ZPT methods; 3) revisit document-level NMT models.

4.1 Experimental Setup

Data and Models We carefully collected training data for building ZPR, sentence-level MT, document-level MT, and reconstruction-based NMT models, which are illustrated in Table 4. According to the domain between training data and our testsets, our training processes can be categorized into in-domain, out-of-domain and mix-domain. As seen, ZPR models of Mov. Subtitle and Q&A forum are trained on their in-domain data, respectively. We concatenate these two datasets as an out-of-domain dataset for the other three domains. For the sentence-level MT model, the Mov. Subtitle and the other domains are trained on the benchmarks from 12.8M OpenSubtitles2018 and 42.8M WMT2021 News, respectively. For the Q&A Forum domain, we further add pseudo-in-domain data generated via a technique of forward-translation (FT) and data augmentation. We also apply the FT technique to the Web Fiction domain. The pseudo-in-domain data of these two domains are also used to tune the document-level models, respectively. For the document-level models, we directly tune the Mov. Subtitle model on Opensubtitle dataset which contains document boundaries. Since the WMT2021 dataset did not contain document boundaries, we follow the setting of (Li et al., 2020) by feeding the model with pseudo sentences

as context. Since there is only Tvsub data containing weak labels, we then tune the model back to its conventional domain by freezing the embedding and reconstruction-related parameters (detailed in Appendix §A.1).

All data are tokenized and then segmented into subword units using the byte-pair encoding (BPE) (Sennrich et al., 2016). We apply 32K merge operations to form a vocabulary for Chinese and English, and the vocabulary is not shared among source and target languages. The proposed dataset has been randomly split into validation and testsets. All models are implemented on top of Transformer (Vaswani et al., 2017a), of which configurations are detailed in Section 4.2. We employed *large-batch training* (i.e. 458K tokens/batch) to optimize the performance (Ott et al., 2018).

Evaluation Metrics We used case-insensitive tokenBLEU (Papineni et al., 2002) to measure the overall translation quality of translation systems,[7] and used *sign-test* (Collins et al., 2005) for testing statistical significance.

To measure the performance of ZPT, we used a variant of APT (Werlen and Popescu-Belis, 2017), which is originally designed to measure the accuracy of explicit pronoun translation. Specifically, the variant AZPT evaluates the accuracy of translating zero pronouns in the source sentences and is calculated by

$$\text{AZPT} = \frac{\sum_{z \in \mathbf{ZP}} A(t_z|z)}{|\mathbf{ZP}|} \quad (1)$$

where **ZP** is the list of zero pronouns in the source sentences, t_z is the generated translation for the zero pronoun z, and $A(t_z|z)$ is a binary scorer to judge whether t_z is the correct translation of z.

In this work, we implemented an automatic version of AZPT for the proposed Chinese-English mZPRT benchmark. For automatically identifying t_z, we obtain the word alignment between ZP-labeled input and its translation output with a GIZA++ model (Och and Ney, 2003), which is trained on a large-scale parallel corpus. To remedy the errors of automatic alignment, for each labeled ZP z, we use the aligned target word along with its preceding and following words as the candidates of t_z. It is not straightforward to implement $A(t_z|z)$, since the Chinese and English pronouns are not one-to-one equivalent. For example, the Chinese pronoun "我" corresponds to two English pronouns (i.e., "I" and "me"). We disambiguate them using pronoun form labels (e.g., subjective, objective in Section 3.2) and a bilingual pronoun dictionary (defined in Appendix §A.2). For example, if the label of Chinese pronoun "我" is *subjective*, its correct translation in English is "I". The automatic implementation of AZPT shows a high correlation (as shown in Table 11, 0.74 Pearson score, 2750 out of 3000 translation annotations reach an agreement) with human judges on the Chinese-English testset, indicating that it is a reasonable metric to evaluate the accuracy of ZPT (Rei et al., 2020; Wan et al., 2022). There are many possible ways to implement the general idea of measuring ZP translation accuracy. The aim of this paper is not to explore the whole space but simply to show that one fairly straightforward implementation works well. The limitation of AZPT is that it can only evaluate Chinese-English ZPT, and we discuss the extension to other languages in Appendix §A.2.

[7] github.com/moses-smt/mosesdecoder/scripts/generic/multi-bleu.perl

4.2 Effects of ZPR on ZPT

As shown in Table 5, we investigate the effects of ZPR models on the ZPT task by studying a pipeline-based method ZPR↦MT (ZPR+), where input sentences are labeled by a ZPR model and then fed into an MT model.

- *MT* (BIG): We built a Transformer model with the BIG settings in (Vaswani et al., 2017a).
- *ZPR*: We followed Song et al. (2020) to build a ZPR model, where BERT (Devlin et al., 2019) is used to represent each input sentence to provide shared features. In practice, we finetune BERT with only ZPR signals instead of jointly learning ZP resolution.

Previous works evaluated the accuracy of ZPR on sentence-level in terms of recall, precision, and F-measure (Wang et al., 2016; Tan et al., 2021; Song et al., 2020). The overall recall and precision on the test set are computed by micro-averaging overall test instances and then the overall F-measure is computed. In our preliminary experiments, we found that the precision of ZPR systems is more impactful to the downstream translation task (e.g. recovering a wrong ZP is more harmful than doing nothing). Thus, we report the precision of recovering ZPs (AZPR) in following experiments.

ZPR Benefits ZPT Considering all ZP types (in the "All" row), most ZPR models can help to improve plain NMT models in terms of ZP translation

Type	Model	Mov. Subtitle		Q&A Forum		Web Fiction		Gov. News		Per. Profile	
		ΔZPR	ΔZPT	ΔZPR	ΔZPT	ΔZPR	ΔZPT	ΔZPR	ΔZPT	ΔZPR	ΔZPT
Sub.	BIG		52.4		26.5		31.5		51.3		62.1
	ZPR+	39.4	↓49.2	61.2	55.8	43.1	33.8	30.5	↓49.4	52.3	63.7
Obj.	BIG		43.5		32.3		18.9		0		0
	ZPR+	52.6	53.0	46.2	45.2	60.0	18.9	n/a	0	n/a	0
Pos.	BIG		30.2		0		24.8		19.9		18.0
	ZPR+	50.9	45.3	74.2	47.2	59.5	31.4	0	↓19.3	41.2	23.9
All	BIG		47.4		25.4		30.9		45.4		58.9
	ZPR+	42.5	49.5	61.5	56.4	46.2	32.0	29.5	↓44.7	50.7	59.7

Table 5: Effects of external ZPR systems on the ZPT task. The baselines are Transformer-BIG models. ZPR+ indicates input sentences are labeled by a ZPR model and then fed into BIG. We report performances of ZPR and ZPT by calculating their accuracy, respectively. 'n/a' and '0' are caused by low frequency and label sparsity.

1. Out-of-Domain	INP.	[他的]$_p$ 主要 研究 领域 为 ...
	BIG	The main research areas are ...
	ZPR	我 主要 研究 领域 为 ...
	ZPR+	My main research areas are ...
2. Error Propagation	INP.	如果 [你们]$_s$ 见到 她 ...
	BIG	If you see her ...
	ZPR	如果 我 见到 她 ...
	ZPR+	If I see her ...
3. Multiple ZPs	INP.	[他]$_s$ 好久没 ... [他]$_s$ 怪 想念 的。
	BIG	for a long time did not ... strange miss.
	ZPR	我 好久没 ... 我 怪 想念 的。
	ZPR+	I haven't ... for a long time, I miss.

Table 6: Errors in ZPR and MT tasks which correspond to Table 5. INP. represents Chinese input and ZPR denotes ZP-annotated output predicted by ZPR models. [Words] are ZPs that are invisible in decoding.

(ΔZPT ↑). Specifically, ZPR+ models can achieve 2∼30% ΔZPT improvements over Transformer-BIG models across different domains. Besides, the higher performance of ZPR models (ΔZPR ↑), the better translation quality of ZPs (ΔZPT ↑). As seen, the ZPR model with 43% accuracy can only achieve +2% ΔZPT points while that with 62% ΔZPR can surprisingly obtain +30% ΔZPT (i.e. movie subtitle vs. Q&A forum). On the contrary, ZPR systems with low-quality (i.e. $<$ 40% ΔZPR points, empirically) harm ZPT quality. Taking government news for instance (out-of-domain ZPR and in-domain MT), the ΔZPT scores decrease by -0.7% when using a ZPR system with only 30% accuracy. The findings are similar in fine-grained cases although there are still considerable differences among different domains.

Large Space for Improvement The best ZPR model (Q&A forum) can only achieve around 62% accuracy, leading to a 56% score on translating ZPs. There is still a large space for further improvement of ZPR and we then identify challenges from three perspectives (case study in Table 6). (1) *out-of-domain*, where it lacks in-domain data for training robust ZPR models. Taking personal profile as an example, the distribution of ZP types is quite different between ZPR training data (out-of-domain) and ZPT testset (in-domain). This leads to that the ZPR model often predicts wrong ZP forms (possessive adjective vs. subject). (2) *error propagation*, where the external ZPR model may provide incorrect ZP words to the followed NMT model. As seen, ZPR+ performs worse than a plain NMT model BIG due to wrong pronouns predicted by the ZPR model (你们 vs. 我). (3) *multiple ZPs*, where there is a 10% percentage of sentences that contain more than two ZPs (as shown in Table 3), resulting in more challenges to accurately and simultaneously predict them. As seen, two ZPs are incorrectly predicted into "我" instead of "他".

4.3 Revisiting NMT Variants

Table 7 shows translation quality (BLEU) and accuracy of ZPT (ΔZPT) across different domains. We investigate three competitive NMT models, three representative ZPT approaches and two oracle methods, as follows (apart from ZPR+):

- *Scaled Transformer*: We trained BASE and BIG Transformer models (Vaswani et al., 2017b). The DEEP model contains 12-12 layers based on

Model	Size	Mov. Subtitle		Q&A Forum		Web Fiction		Gov. News		Per. Profile	
		BLEU	AZPT	BLEU	AZPT	BLEU	AZPT	BLEU	AZPT	BLEU	AZPT
Scaled Transformer											
BIG	301M	29.4	47.4	12.7	25.4	11.7	30.9	21.0	45.4	39.2	58.9
BASE	106M	28.3	50.1	11.9	29.0	11.4	31.5	19.1	41.5	37.1	53.4
DEEP	477M	30.0	47.4	12.8	30.2	11.7	31.8	21.5	42.4	41.2	58.5
Existing ZPT Methods (on top of BIG*)*											
ZPR+	301M	29.8	49.5	13.2	56.4	11.6	32.0	20.9	44.6	38.9	59.7
DOC.	301M	29.8	53.5	13.9	26.3	12.2	35.3	20.5	46.1	38.7	59.3
REC.	453M	30.0	52.3	12.3	30.4	12.0	33.4	21.0	46.3	38.3	60.6
Oracle											
DEEP	477M	32.8	86.9	14.7	88.8	12.8	85.1	21.7	89.0	41.5	90.3
REC.	453M	33.1	89.7	14.2	89.2	12.5	86.1	21.4	89.1	40.9	91.3

Table 7: A benchmark of ZPT was evaluated on the proposed dataset. We report overall translation quality and translation accuracy of ZPs with BLEU and AZPT. Darker color denotes more improvement over the Transformer-BIG. Most colored values are statistically significant difference ($p < 0.01$) from the BIG model.

BASE configurations.

- *Document-Level NMT* (DOC.): We used the unified encoder (Ma et al., 2020), which takes the concatenation of contexts and source sentences as the input using two-level self-attention. Instead of using BERT pretraining, we first train the Transformer model on sentence-level training data until converged and then switch to document-level training data (Zhang et al., 2018).
- *Reconstruction-based NMT* (REC.): We re-implemented the model provided by Wang et al. (2018a), where two additional reconstructors (Tu et al., 2017) are introduced to reconstruct ZP-labeled source sentences for both encoder and decoder. The auxiliary training objectives can encourage the latent representations to embed ZP information.
- *Oracle*: We manually annotated ZPs in input sentences and then feed them into downstream MT/ZPT models. This can be regarded as the "upper bound" performance the models can reach.

Scaling Transformer Cannot Benefit ZPT Although the overall translation quality is significantly improved by scaling models (Size ↑ vs. BLEU ↑), the accuracy of translating ZPs still can not be guaranteed (AZPT). For example, BIG and DEEP with larger parameters can achieve better performance than the BASE in terms of BLEU (+1.2 and +1.9 points on average). However, the corresponding AZPT scores remain almost unchanged or slightly declined (+0.5 point on average).

Existing Methods Can Help ZPT But Not Enough Three ZPT models can improve ZP translation in most cases, although there are still considerable differences among different domains (AZPT ↑). Introducing ZPT methods has little impact on BLEU score (-0.1∼+0.2 point on average), however, they can improve AZPT over BIG by +2.2∼+6.8. When integrating golden ZP labels into DEEP and REC. models, their BLEU and AZPT scores largely increased by +1.9 and +47.5 points, respectively. The performance gap between Oracle and others shows that there is still a large space for further improvement for ZPT.

Evaluation on Diagnostic Subset Someone may argue that resolving ZP can not significantly improve BLEU scores. We establish a benchmark on diagnostic subsets (only ZP sentences) that accounts for an average of 44% of the full testset. As shown in Table 8, the ZPT approaches are showing more gains in translation quality, which indicates the necessity of recovering ZPs. Compare to the result in Table 7, using oracle sequence as input achieves larger improvement on diagnostic subsets (+2.1 v.s. +1.1 BLEU on average), especially on the in-domain dataset (over 4.0 BLEU on average). This demonstrates the importance of recovering the ZPs on translation tasks.

4.4 Revisiting Document-Level NMT

One commonly-cited weakness in document-level NMT is that general-purpose metrics (e.g. BLEU) are not sufficient to distinguish translation qualities

Model	M.S.	Q.F.	W.F.	G.N.	P.P.
Scaled Transformer					
BIG	29.5	12.4	10.7	19.9	39.5
BASE	27.8	11.4	10.2	18.2	37.4
DEEP	28.6	11.8	10.9	20.3	41.3
Existing ZPT Methods (on top of BIG*)*					
ZPR+	30.7	13.2	12.2	19.8	39.1
DOC.	31.1	13.7	11.9	20.1	39.2
REC.	30.3	12.5	11.2	20.0	39.1
Oracle					
DEEP	34.9	14.1	12.6	20.5	41.7
REC.	35.1	14.0	12.5	20.3	41.4

Table 8: Evaluation on diagnostic subsets, where we extract ZP sentences from the proposed dataset. We only report the BLEU scores because the AZPT scores are the same as Table 7. M.S., Q.F., W.F., G.N., P.P. are the abbreviations of Mov. Subtitle, Q&A Forum, Web Fiction, Gov. News, and Per. Profile.

Model	BLEU	AZPT	Human
BIG	29.4	47.4	3.47
MULTI-ENC	28.6	48.6	3.50
UNIFIED	29.8	53.5	3.96
CACHE	29.4	51.2	3.57

Table 9: Evaluation of document-level NMT models on movie subtitle using BLEU, AZPT and human scores.

from the perspective of discourse (Müller et al., 2018; Voita et al., 2018, 2019; Xu et al., 2021). As ZP is a significant phenomenon of cohesion, the proposed dataset and metric are complementary to verify different document-level approaches. As shown in Table 9, we revisit three advanced models: multi-encoder (Zhang et al., 2018), unified encoder (Ma et al., 2020) and cache-based (Tu et al., 2018). Encouragingly, we find that the trend of AZPT scores is more close to true performance. For example, the CACHE model has no improvement in terms of BLEU while it can achieve +3.8% better performance on translating ZPs.

5 Conclusion

We revealed data- and evaluation-level gaps in previous works on translating ZPs. Accordingly, we proposed a benchmark testset and evaluation metric for target evaluation on ZPT. Our benchmark emphasizes the large gap between performances of existing models and an upper bound from the perspective of ZPT. We release data (benchmark testset, human judgments and collected training data), code (target metric and implemented approaches) and models (comparative models), which we hope can significantly promote research in this field.

Limitation

We list the main limitations of this work as follows:

1. *Extending The Dataset to Other Languages*: The zero pronoun (ZP) phenomenon may vary across languages in terms of word form, occurrence frequency and category distribution etc. This work mainly proposed a ZP translation testset in Chinese-English due to the lack of linguistic experts in other languages. Another reason is the high cost of human annotation, where we spent $10,000 US dollars for annotating 8,093 sentences in five domains. The annotation guideline in Section 3.2 could be adapted to other languages with the corresponding linguistic experts. Therefore, one aim of our work is attracting more attention to this research topic, which will stimulate further contributions on building multilingual resources.
2. *The* AZPT *Metric*: We used the variant of APT (Werlen and Popescu-Belis, 2017) as a target evaluation metric for the proposed Chinese-English benchmark. This metric may not be applicable to all languages, because the situation of one-to-many pronoun gender is not considered. In Appendix §A.2, we have revealed the reason behind, and provided alternative ways to extend it to other languages. Besides, the reliability of such metrics depend on the quality of word alignment models. We trained the alignment model using GIZA++ on a large-scale parallel corpus, which still has 10% to 15% deviation between automatic and human evaluation. This gap could be narrowed by further improving the alignment quality. In the paper, we do not list the AZPT metric as one of our main contributions due to its potential limitation on applicability to other languages.

Acknowledgements

This work was supported in part by the Science and Technology Development Fund, Macau SAR (Grant No. 0101/2019/A2), and the Multi-year Research Grant from the University of Macau (Grant No. MYRG2020-00054-FST).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Elizabeth Baran, Yaqin Yang, and Nianwen Xue. 2012. Annotating dropped pronouns in chinese newswire text. In *LREC 2012*.

Tagyoung Chung and Daniel Gildea. 2010. Effects of empty categories on machine translation. In *EMNLP*.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *ACL*.

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

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML*.

Liane Guillou and Christian Hardmeier. 2018. Automatic reference-based evaluation of pronoun translation misses the point. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4797–4802.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *IWSLT*.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *ACL*.

Charles N Li and Sandra A Thompson. 1979. Third-person pronouns and zero-anaphora in chinese discourse. In *Discourse and syntax*. Brill.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *ACL*.

Aya A Mitani, Phoebe E Freer, and Kerrie P Nelson. 2017. Summary measures of agreement and association between many raters' ordinal classifications. *Annals of epidemiology*.

Ruslan Mitkov. 2014. *Anaphora resolution*. Routledge.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *WMT*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. In *CL*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL*.

Jesús Peral and Antonio Ferrández. 2003. Translation of pronominal anaphora between english and spanish: Discrepancies and evaluation. In *JAIR*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *CoNLL-WS*.

Sudha Rao, Allyson Ettinger, Hal Daumé III, and Philip Resnik. 2015. Dialogue focus tracking for zero pronoun resolution. In *NAACL*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *EMNLP*.

Ryokan Ri, Toshiaki Nakazawa, and Yoshimasa Tsuruoka. 2021. Zero-pronoun data augmentation for japanese-to-english translation. In *WAT*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL*.

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 *ACL*.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving Multi-turn Dialogue Modelling with Utterance ReWriter. In *ACL*.

Xin Tan, Longyin Zhang, and Guodong Zhou. 2021. Coupling context modeling with zero pronoun recovering for document-level natural language generation. In *EMNLP*.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *AAAI*.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. In *TACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *NeurIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *NIPS*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *ACL*.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *ACL*.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *ACL*.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018a. Translating pro-drop languages with reconstruction models. In *AAAI*.

Longyue Wang, Zhaopeng Tu, Xing Wang, and Shuming Shi. 2019. One model to learn both: Zero pronoun prediction and translation. In *EMNLP-IJCNLP*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017a. Exploiting cross-sentence context for neural machine translation. In *EMNLP*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2018b. Learning to jointly translate and predict dropped pronouns with a shared reconstruction mechanism. In *EMNLP*.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016. A novel approach to dropped pronoun translation. In *NAACL*.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Siyou Liu, Hang Li, Andy Way, and Qun Liu. 2017b. A novel and robust approach for pro-drop language translation. *Machine Translation*, 31(1):65–87.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (apt). In *DiscoMT*.

Mingzhou Xu, Liangyou Li, Derek F. Wong, Qun Liu, and Lidia S. Chao. 2021. Document graph for neural machine translation. In *EMNLP*.

Jingxuan Yang, Jianzhuo Tong, Si Li, Sheng Gao, Jun Guo, and Nianwen Xue. 2019. Recovering dropped pronouns in Chinese conversations via modeling their referents. In *NAACL*.

Yaqin Yang, Yalin Liu, and Nianwen Xue. 2015. Recovering dropped pronouns from chinese text messages. In *ACL-IJCNLP*.

Yaqin Yang and Nianwen Xue. 2010. Chasing the ghost: recovering empty categories in the chinese treebank. In *COLING*.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with bayes' rule. In *TACL*.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *EMNLP*.

Weinan Zhang, Ting Liu, Qingyu Yin, and Yu Zhang. 2019. Neural recovery machine for Chinese dropped pronoun. In *Frontiers of Computer Science*.

A Appendix

A.1 Experimental Setup

Machine Translation We build several domain-specific translation models to match the domains of the proposed testing dataset. To train NMT models in movie subtitle and government news domains, we used 12.8M OpenSubtitles2018[8] and 42.8M WMT2021 News[9] parallel corpora, respectively. For the web fiction domain, we extracted 1.45M Chinese texts from the Webnovel website and then employed the technique of forward-translation (FT) to generate the in-domain data. About the Q&A Forum domain, we reused the 50K training data of Baidu Knows corpus as monolingual data, and also employed FT to get the domain-specific training data. Due to an insufficient amount of in-domain monolingual data, we used ten commercial translation systems to construct 10×5K synthetic parallel data. Note that since the scale of these pseudo data is much less than WMT2021, we tuned the baseline model on the pseudo data to get the domain-specific variant directly in these two domains. We train an NMT model on WMT2021 News for the Personal Profile domain because we regard it as an out-of-domain issue.

Zero Pronoun Recovery and Translation The ZPR systems can be used to pre-process input sentences before being fed into NMT models, namely the pipeline-based ZPT method (Table 7 ZPR+). We used 2.2M TVsub (Wang et al., 2018a) and 5K BaiduKnows to train two ZPR models for movie subtitle and Q&A Forum testsets, respectively. For the other three domains, we combine these two corpora to train a general-domain ZPR model. Accordingly, document-level NMT models (Table 7 DOC.) are trained on several domains of datasets. For movie subtitle, web fiction and Q&A forum domains, we used the same data used in corresponding baselines with their context and document boundary information. Regarding government news and personal profile domains, there are no context-aware corpora. Thus, we feed the model with pseudo sentences as context, where context representations act more like a noise generator to provide richer training signals (Li et al., 2020). Reconstruction-based NMT models (Table 7 REC.) are pre-trained on the same data used in corresponding baselines and then tune reconstruction-related parameters with only monolingual ZP datasets. Since there is only Tvsub data containing weak labels, we then tune the model back to its conventional domain by freezing the embedding and reconstruction-related parameters.

[8] https://opus.nlpl.eu/OpenSubtitles-v2018.php.

[9] http://www.statmt.org/wmt21/translation-task.html.

A.2 Discussion on AZPT

Extensibility As shown in Table 10, Chinese-English pronouns are many-to-many mapped. In fact, pronoun-aware evaluation metrics mainly focus on solving one-to-many problems:

- *pronoun form*: to disambiguate multiple pronoun translations according to the sentence unit of source ZPs. Taking 我 ⇒ I/me for example, it should be translated into "I" when "我" is subject while "me" when object. Therefore, we annotate our dataset with fine-grained form labels (in Section 3.2) and use this information in AZPT (in Section 4.1).
- *pronoun gender*: to disambiguate multiple pronoun translations according to the antecedent gender of source ZPs. However, this is not a problem for Chinese-English due to many-to-one mapping (e.g. 他们/她们/它们 ⇒ them).

However, the *pronoun gender* is a problem for other pro-drop languages (e.g. French and Spanish) due to one-to-many mapping. For example, *Audi is an automaker that makes luxury cars. It was established by August Horch.* The pronoun "it" could be translated into "il" (masculine singular subject pronoun), "elle" (feminine singular subject pronoun) or "cela" (demonstrative pronoun) according its antecedent gender "Audi". Here we provide two alternative ways to extend AZPT: (1) adding fine-grained gender labels to ZPs, which is similar to AutoPRF (Hardmeier and Federico, 2010); (2) introducing translation reference as soft information, which is similar to APT (Werlen and Popescu-Belis, 2017).

Human Evaluation Guideline We carefully design an evaluation protocol according to error types made by various NMT systems, which can be grouped into five categories: 1) The translation can not preserve the original semantics due to misunderstanding the anaphora of ZPs. Furthermore, the structure of translation is inappropriately or grammatically incorrect due to incorrect ZPs or lack of ZPs; 2) The sentence structure is correct, but translation can not preserve the original semantics

Form	Subject	Object	Possessive adjective	Possessive	Reflexive
1st SG	我 (I)	我 (me)	我的 (my)	我的 (mine)	我自己的 (myself)
2nd SG	你 (you)	你 (you)	你的 (your)	你的 (yours)	你自己的 (yourself)
3rd SGM	他 (he)	他 (him)	他的 (his)	他的 (his)	他自己的 (himself)
3rd SGF	她 (she)	她 (her)	她的 (her)	她的 (hers)	她自己的 (herself)
3rd SGN	它 (it)	它 (me)	它的 (its)	它的 (its)	它自己的 (itself)
1st PL	我们 (we)	我们 (us)	你们的 (your)	你们的 (yours)	你们自己的 (yourselves)
2nd PL	你们 (you)	你们 (you)	我们的 (our)	我们的 (ours)	我们自己的 (ourselves)
3rd PLM	他们 (they)	他们 (them)	他们的 (their)	他们的 (theirs)	他们自己的 (themselves)
3rd PLF	她们 (they)	她们 (them)	她们的 (their)	她们的 (theirs)	她们自己的 (themselves)
3rd PLN	它们 (they)	它们 (them)	它们的 (their)	它们的 (theirs)	它们自己的 (themselves)

Table 10: Chinese-English pronouns with corresponding forms. The pronoun types are short for: person = 1st, 2nd, 3rd, singular = SG, plural = PL, male = M, female = F and neutral = N.

	AZPT	APT	BLEU	TER	MET.	COM.
M.S.	**0.68**	0.56	0.09	0.41	0.23	0.59
W.F.	0.73	0.42	0.50	0.22	0.67	**0.80**
Q.F.	0.76	**0.84**	0.38	0.01	0.74	0.15
G.N.	**0.58**	0.25	0.57	0.26	0.28	0.37
P.P.	**0.96**	0.54	0.62	0.68	0.59	0.71
AVE.	**0.74**	0.52	0.43	0.32	0.38	0.52

Table 11: Correlation between the manual evaluation and other automatic metrics, which are applied to different domains of the mZPRT dataset. MET. denotes Meteor metric and COM. is COMET, a neural-based automatic metric.

due to misunderstanding the anaphora of ZPs; 3) The translation can preserve the original semantics, but the structure of translation is inappropriately generated or grammatically incorrect due to the lack of ZPs; 4) where a source ZP is incorrectly translated or not translated, but the translation can reflect the meaning of the source; 5) where translation preserves the meaning of the source and all ZPs are translated. Finally, we average the score of each target sentence that contains ZPs to be the final score of our human evaluation.

For human evaluation, we randomly select a hundred groups of samples from each domain, each group contains an oracle source sentence and the hypotheses from six examined MT systems. Following this protocol, we asked expert raters to score all of these samples in 1 to 5 scores to reflect the quality of ZP translations. As shown in Table 11, our variant AZPT reaches around 0.74 Pearson scores with human judges, while APT reaches only 0.52. This confirms that the AZPT takes great success in adapting to Chinese to English ZPT task. For the inter-agreement, we simply define that a large than 3 is a good translation and a bad translation is less than 3. The annotators reached an agreement of annotations on 91% (2750 out of 3000) samples. Considering domain-specific subsets, AZPT achieves the best scores on three domains and the second on others, among automatic metrics. The web fiction domain contains a number of free translations due to its literariness. We found that COMET performs better in evaluating semantic correctness by leveraging cross-lingual pre-training. In the Q&A forum subset, sentences are shorter but contain more multiple ZPs (in Table 3). We observe that reference is helpful to APT to disambiguate neighbor ZPs in one sentence.