Coupling Context Modeling with Zero Pronoun Recovering for Document-Level Natural Language Generation

Xin Tan, Longyin Zhang, and Guodong Zhou*
School of Computer Science and Technology, Soochow University, China
{xtan9,lyzhang9}@stu.suda.edu.cn
gdzhou@suda.edu.cn

Abstract

Natural language generation (NLG) tasks on pro-drop languages are known to suffer from zero pronoun (ZP) problems, and the problems remain challenging due to the scarcity of ZP-annotated NLG corpora. In this case, we propose a highly adaptive two-stage approach to couple context modeling with ZP recovering to mitigate the ZP problem in NLG tasks. Notably, we frame the recovery process in a task-supervised fashion where the ZP representation recovering capability is learned during the NLG task learning process, thus our method does not require NLG corpora annotated with ZPs. For system enhancement, we learn an adversarial bot to adjust our model outputs to alleviate the error propagation caused by mis-recovered ZPs. Experiments on three document-level NLG tasks, i.e., machine translation, question answering, and summarization, show that our approach can improve the performance to a great extent, and the improvement on pronoun translation is very impressive.

1 Introduction

For a long time, natural language generation (NLG) has attracted a lot of attention for its importance in serving human life. In the literature, although various studies have been done to bridge the discrepancy between human and machine, document-level NLG (D-NLG) tasks still suffer from cohesion issues caused by zero pronoun (ZP). As a discourse phenomenon where pronouns can be omitted when they are pragmatically or grammatically inferable from context (Li and Thompson, 1979), zero pronoun appears frequently in pro-drop languages like Chinese, Spanish, etc. Taking the Chinese TED corpus as an example, according to our statistics, in sentences with an average length of 18, each sentence will omit around 0.5 pronouns. Facing this problem, lots of attention has been paid to ZP resolution in the past decade (Zhao and Ng, 2007; Kong and Zhou, 2010; Yin et al., 2018; Zhang et al., 2019a; Song et al., 2020), and these studies have achieved certain success. Nevertheless, due to the lack of NLG corpora annotated with ZPs, the problem of zero pronoun remains challenging in document-level NLG tasks.

Recently, more and more researchers turn to ZP resolution in NLG tasks (Rao et al., 2015; Wang et al., 2016, 2018a,b, 2019). Among these studies, one line explicitly deals with the ZP problem by recovering dropped pronouns through either annotated corpora (Wang et al., 2016, 2018a,b, 2019; Zhang et al., 2019c) or pre-trained ZP resolution systems (Taira et al., 2012; Xiang et al., 2013). Another line indirectly deals with the ZP problem by producing better discourse cohesion through document context modeling (Zhang et al., 2018; Miculicich et al., 2018; Tan et al., 2019; Wong et al., 2020). Although the above studies have made great progresses, ZP resolution in NLG still faces the following possible bottlenecks: (i) ZP-annotated corpora tailored for NLG tasks are scarcity, and existing ZP corpora are limited to certain domains and tasks. (ii) Using pre-trained ZP resolution systems to recover pronoun labels for NLG tasks could lead to the notorious error propagation problem. (iii) Although document context modeling can improve discourse-level cohesion to some extent, the context information is too broad to solve the ZP problem in a targeted manner.

In order to solve the ZP problem in D-NLG tasks while avoiding the above disadvantages, we introduce a highly adaptive two-stage approach that couples context modeling with ZP recovering. Specifically, our approach mainly consists of two phases: First, we pre-train a fault-tolerant ZP position detector for downstream tasks' data corpora to automatically detect ZP positions. Second, we perform document context modeling for both task-supervised ZP recovering and ZP-focused NLG task learning. Notably, the ZP recovering process

*Corresponding author

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2530–2540
November 7–11, 2021.

and the NLG task learning process depend on and promote with each other harmoniously. On the one hand, instead of recovering specific pronoun labels, we learn ZP representation through the supervision of NLG tasks and thus our method does not require large-scale ZP-annotated data. One the other, the achieved ZP representation is further fused into the previously modeled document context for ZP-focused NLG task learning in turn. In this way, the recovered ZP representation and the supervision of specific tasks are well shared between the two processes for high-quality model integration.

To comprehensively investigate our proposed method, we conduct experiments[1] on three D-NLG tasks: document-level neural machine translation (NMT), question answering (QA), and summarization. Experimental results show that our approach can significantly improve the performance on these tasks due to the effective combination of ZP recovering and document context modeling. Furthermore, we use both APT (Miculicich Werlen and Popescu-Belis, 2017) and CRC (Jwalapuram et al., 2019) to evaluate our model performance on pronoun generation, and the results show that our approach can achieve remarkable performance.

2 Related Work

As a fundamental research in natural language processing, ZP resolution aims at detecting pronoun chains and resolving missing pronouns to their antecedents. In the literature, previous work mainly resolved ZPs in three steps: zero pronoun detection, anaphoricity determination, and coreference linking. On this basis, varied traditional rule-based or machine-learning methods were used for ZP resolution (Converse, 2006; Zhao and Ng, 2007; Kong and Zhou, 2010). Recently, some neural approaches (Liu et al., 2017; Yin et al., 2018; Zhang et al., 2019a,b; Song et al., 2020) were proposed and have achieved certain success due to their better objective representation and powerful neural architectures.

As a common language phenomenon in pro-drop languages, zero pronoun could result in poor discourse-level cohesion and thus seriously impact the performance of document-level NLG. To date, two types of researches have been conducted to alleviate the discourse-level cohesion deficiency: (i) Recovering dropped pronouns in specific NLG corpora for downstream tasks; (ii) Using well-designed context-aware architectures for document-level cohesion modeling. **First**, some studies directly used manually (Yang et al., 2015; Zhang et al., 2019c; Yang et al., 2020) or automatically (Wang et al., 2016, 2018a) annotated ZP corpora for NLG tasks. Nevertheless, the manual annotation is usually time consuming and the automatic annotation is limited to specific tasks like machine translation, it remains challenging when facing new corpus domains or tasks. Moreover, although some two-stage methods were proposed to use pre-trained ZP resolution systems for pre-processing (Taira et al., 2012; Xiang et al., 2013), these methods are known to face notorious error propagation problems. **Second**, some recent studies explored context-aware architectures for better document cohesion modeling (Zhang et al., 2018; Miculicich et al., 2018; Maruf and Haffari, 2018; Maruf et al., 2019; Tan et al., 2019; Kang et al., 2020). For instance, Tan et al. (2019) proposed a hierarchical model to capture global context, which can significantly improve pronoun translation in document-level NMT. Although the above studies can well capture document-level cohesion to some extent, the context information is too broad to solve ZP problems in a targeted manner.

The difference between our method and previous ones is two-fold: First, our two-stage method is a combination of the above categories which can mitigate both corpora limitation and error propagation issues. Second, compared with previous ZP recovery methods, we focus on enhancing NLG with the recovered ZP representations rather than with specific pronoun labels. Therefore, our approach does not require ZP-annotated NLG corpora.

3 D-NLG with ZP Recovery

In this section, we introduce the proposed highly adaptive approach which consists of two stages, i.e., detecting ZP positions in the first stage (Section 3.1) and then coupling context modeling with ZP representation recovering in document-level NLG (Sections 3.2 and 3.3).

3.1 ZP Position Detection

Due to corpus limitation, previous two-stage methods usually employ pre-trained ZP detectors to automatically recover dropped pronouns for NLG tasks. However, these methods usually suffer from error propagation problems. In addition, referring

[1]The codes (PyTorch) will be published at https://github.com/txAnnie/ZP-DNLG.

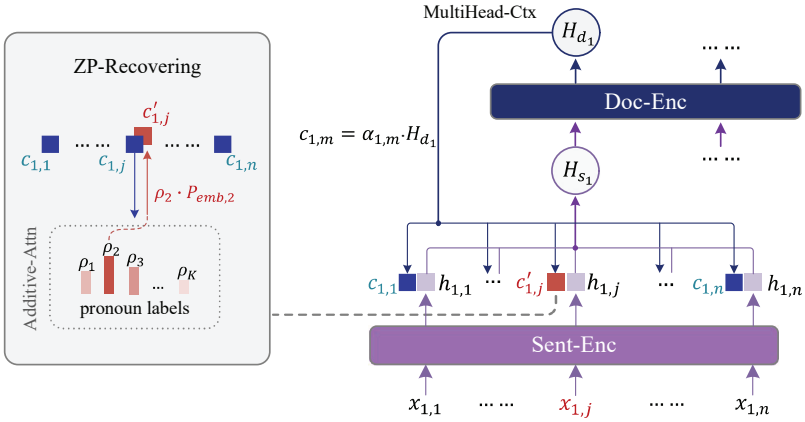

Figure 1: The overall model architecture of our approach. The unit $x_{1,j}$ colored in red refers to the embedded placeholder and $P_{emb,\iota}$ denotes the representation of the ι-th pronoun label.

to the human annotation (Yang et al., 2015), we find that many annotated pronouns are replaceable from the sentence-level perspective, while are irreplaceable from the document-level perspective. With this in mind, we argue that only detecting ZP positions in the first stage and then performing context-aware ZP recovery in the second stage can alleviate the above problems to a certain extent. Notably, considering that ZP position detection is less affected by domain differences, we only need small-scale out-of-domain ZP-annotated data instead of the one that is specific to the target task to achieve it.

In this work, we cast ZP position detection as a sequence labelling task. Our statistics show that in 10% of the cases, the last word of a sentence is a ZP position, which means taking both the left and right sides of each word as candidate ZP positions is necessary. Therefore, different from previous work (Wang et al., 2016), we take both sides of a word into consideration for ZP position detection. Formally, given a sentence with n word units, $w_0, \ldots, w_{n-1}$, the components of the detector can be described as following:

Encoder. A two-layer bi-directional GRU (Cho et al., 2014) is used to map the word units into a set of hidden states $u_0, \ldots, u_{n-1}$.

Decoder. During decoding, we input the previously obtained hidden states into a uni-directional GRU for ZP position prediction. Since both the left and right sides of each word could be ZP positions, we insert placeholders to both sides of each word as candidate ZP positions. Therefore, the decoder input can be formulated as $(\epsilon, u_0, \ldots, \epsilon, u_{n-1}, \epsilon)$ where the sign ϵ denotes a randomly initialized vector. It should be noted that although these placeholders share the same learnable vector ϵ, the corresponding decoder outputs of these placeholders are context-aware and definitely different.

Based on the model structure, we simply build a negative log-likelihood loss term between the decoder outputs (after log-softmax) and the gold standard ZP positions to train our ZP position detector. It is worth mentioning that we aim to build a fault-tolerant ZP position detector, in other words, we permit mis-predicted ZP positions to appear in this stage and this will provide more possibilities and rights for subsequent tasks (Section 3.2) to further determine the value of these positions.

3.2 Context Modeling for ZP Recovering

In this work, we hold the view that the recovery of ZPs needs not only sentence-level semantics or intentions but also document-level context. With this in mind, based on the previously detected ZP positions, we explore to leverage document context for ZP representation recovery in this section.

Document-level Context Modeling. Hierarchical architecture has proven to be effective in document context modeling in many NLG tasks. Recently, Tan et al. (2019) demonstrate that global document context performs better than the partial one in document-level NMT. Inspired by this, we also employ a hierarchical network to model global context for ZP recovery, as shown in Figure 1. Formally, given a document with N sentences, the

context modeling process is formulated as:

$$h_{s_i} = \text{ENCODER}_{sent}(s_i) \quad (1)$$

$$H_{s_i} = \sum \text{ATT}_{self}(h_{s_i}, h_{s_i}, h_{s_i}) \quad (2)$$

$$H_d = \text{ENCODER}_{doc}(H_{s_1}, \ldots, H_{s_N}) \quad (3)$$

where $s_i = (x_{i,1}, \ldots, x_{i,n})$ denotes the sentence i with n word units, $h_{s_i} = (h_{i,1}, \ldots, h_{i,n})$ denotes the hidden representation of the words in the sentence, $H_{s_i} \in \mathbb{R}^{D\times 1}$ is a weighted representation of the sentence, $H_d = (H_{d_1}, \ldots, H_{d_N}) \in \mathbb{R}^{D\times N}$ denotes the extracted context information, and ATT_{self} denotes a multi-head self-attention function mentioned in (Vaswani et al., 2017). Following Tan et al. (2019), we implement both the sentence- and document-level encoders with multi-head self-attention functions, and the model parameters are shared between the two encoders.

Context-Aware ZP Recovering. Usually, document context consists of not only relationships between sentences, but also dependencies among words. Obviously, each word has its specific surrounding context, even those ZP placeholders, and it is reasonable to utilize the surrounding context of ZP placeholders for ZP representation recovery. To achieve this, inspired by (Tan et al., 2019), we distribute the previously obtained context information to each word unit as:

$$\alpha_{i,j} = \text{ATT}_{additive}(H_{d_i}, h_{i,j}) \quad (4)$$

$$c_{i,j} = \alpha_{i,j} \cdot H_{d_i} \quad (5)$$

where $h_{i,j}$ is the hidden state of the j-th word in the i-th sentence, H_{d_i} denotes the extracted document context, $\text{ATT}_{additive}$ is an additive attention function, $\alpha_{i,j}$ and $c_{i,j}$ denote the attention weight and context information assigned to the word. The complete context modeling and distributing process is clearly illustrated in Figure 1.

Before introducing the ZP representation recovering process, an important question needs to be clarified: Do downstream tasks require all the ZP positions to be recovered? Taking machine translation for example. Given the sentence "(他们) 都 用 了 fMRI 技术 也 就是 功能性 核磁共振 成像 技术 来 对 大脑 进行 造影 。" with the subject "他们 (They)" omitted, the reference translation is "Both used fmri technology functional magnetic resonance imaging technology to image the brain ." Obviously, the dropped pronoun "他们 (They)" is not explicitly translated, according to the language habits of the target side. This indicates that not all the dropped pronouns need to be recovered, and it requires a good understanding of the context to determine whether pronouns like "他们 (They)" should be recovered or not for better translation. Furthermore, since we perform fault-tolerant ZP position detection in the first stage, it will naturally contain some mis-predicted ZP positions. Taking into account the above-mentioned circumstances, we add a non-ZP mark ε in the label space for our model to determine whether the detected ZP positions should be filled with specific pronoun labels or not according to the document context.

For ZP recovery, given the document context assigned to each ZP position, we build another additive attention function between the context information, $c_{i,j}$, assigned to each ZP placeholder and the pronoun label vectors[2] P_{emb}, as shown in Figure 1. The attention scores are calculated as:

$$\rho_1, \ldots, \rho_K = \text{ATT}_{additive}(c_{i,j}, P_{emb}) \quad (6)$$

where $K = 31$ denotes the number of pronoun labels. As stated before, in this work, we propose to borrow the learning objectives from NLG tasks rather than manually annotated ZP corpora to guide the learning of ZP representation recovery. Specifically, for each placeholder, we first select the pronoun label vector with the highest attention score and then multiply the selected vector with its corresponding attention weight ρ_ι as the recovered ZP representation, which can be written as:

$$c'_{i,j} = \max_{\iota}(\rho_1, \ldots, \rho_K) \cdot P_{emb,\iota} \quad (7)$$

In this way, as the gradient is updated, both the ZP representations and the vector-style pronoun label selection process are learned automatically.

3.3 D-NLG with ZP Representation

In this subsection, we aim to integrate the ZP recovery process into specific document-level NLG tasks. The integration process is mainly composed of two phases: First, we replace the original context information distributed to each placeholder, $c_{i,j}$, with the obtained ZP representation, $c'_{i,j}$, for ZP-focused context modeling; Second, we combine the ZP-focused context with the sentence-level encoder outputs and push the combinations into the decoding phase of each subsequent task for the learning of NLG tasks.

[2]The pronoun vectors are randomly initialized 512D vectors that represent the 31 pronoun labels (including the ε label); the pronoun labels are detailed in Appendix.

In detail, we apply our method to three recent NLG systems: document-level NMT (Tan et al., 2019), QA and summarization (Xu et al., 2020). Among them, Tan et al. (2019) introduced a hierarchical structure to model global context from all sentences of an article and have demonstrated the effectiveness of global context in machine translation. Xu et al. (2020) presented a straightforward yet effective model on summarization and QA based on their proposed MATINF dataset[3]. We incorporate our ZP-focused context information into the three systems as following:

- For document-level NMT, we first apply our ZP position detector to NMT corpora for preprocessing. Then, based on the Transformer-based system of Tan et al. (2019), we input the sentences with ZP positions into their encoder for ZP representation recovering and global context refining, and other settings remain the same as theirs.

- For QA and summarization, we also preprocess the MATINF corpus with ZP position detection. Similar to NMT, in QA and summarization, we extract global context for each word unit for both ZP recovering and context refining. Since Xu et al. (2020) did not extract document context in their original system, we simply incorporate the obtained ZP-focused context information into their original word representation through a summation function. And other settings remain the same as (Xu et al., 2020).

3.4 Model Learning

The overall model learning is composed of two parts: (i) Learning from the NLG tasks' objectives for standard language generation; (ii) Training our language generator adversarially to reduce the errors caused by mis-recovered ZPs.

First, we train our model according to the NLG tasks' objectives to warm up the model parameters. That is, we maximize the log-likelihood of language generation in the parallel corpus C as:

$$\hat{\theta} = \arg\max_{\theta} \sum_{<x,y>\in C} \log P(y|x;\theta) \quad (8)$$

After that, the adversarial nets participate in the model learning process. Notably, we build two CNN-based feature extractors for the model outputs $\boldsymbol{o}$ and the references $\boldsymbol{r}$, respectively. The feature extractor is formulated as:

$$\tau_i = \mathrm{F}_{relu}(w \cdot y_{i:i+k} + b) \quad (9)$$

$$\overline{\tau} = \mathrm{mean}(\tau_1, \ldots, \tau_m) \quad (10)$$

where we set the window size by 4 ($k = 3$) and the stride size by 1, and F_{relu} refers to the Leaky ReLU activation function. In this way, the extracted features for $\boldsymbol{o}$ and $\boldsymbol{r}$ can be written as $\overline{\tau}_{\boldsymbol{o}}$ and $\overline{\tau}_{\boldsymbol{r}}$, respectively.

Our adversarial nets consist of two parts: (i) A generative net $G(X, \theta_{\mathbf{g}})$ that builds the mapping from model outputs to feature space to capture the data distribution $p_{\mathbf{g}}$ over the training data X. (ii) A discriminative net $D(\overline{\tau}, \theta_{\mathbf{d}})$ that outputs a single scalar representing the probability that the extracted feature $\overline{\tau}$ comes from training data X rather than $p_{\mathbf{g}}$. On this basis, we let G and D join the training process to play a two-player minimax game. Given the feature of the generated samples, we train G to maximize the following object:

$$\mathcal{L}(\theta_{\mathbf{g}}) = \sum \log(\hat{p}^{\mathbf{g}}_D(\overline{\tau}_{\boldsymbol{o}})) \quad (11)$$

where $\theta_{\mathbf{g}}$ refers to parameters of our NLG model and the feature extractor for the model outputs. After that, we simultaneously train D to maximize the probability of assigning correct labels to both gold standard and fake generated samples. Formally, we train D to minimize the following object:

$$\mathcal{L}(\theta_{\mathbf{d}}) = -\sum \log(\hat{p}^{\mathbf{d}}_D(\overline{\tau}_{\boldsymbol{r}})) - \sum \log(1 - \hat{p}^{\mathbf{d}}_D(\overline{\tau}_{\boldsymbol{o}})) \quad (12)$$

where $\theta_{\mathbf{d}}$ refers to the parameters of the feature extractor for the references and a feedforward network-based (In: feature size f, Hidden layer: $f/2$, Out: 1) scorer with sigmoid function.

4 Experimentation

In this section, we conduct several experiments on document-level NMT, QA, and summarization to evaluate our proposed approach.

4.1 Experimental Settings

Datasets. For document-level NMT, previous studies (Maruf and Haffari, 2018; Maruf et al., 2019; Tan et al., 2019) usually employ a two-stage training strategy for model learning. Following (Tan et al., 2019), we also use a 2.8M sentence-level corpus from news corpora LDC2003E14,

[3]Maternal and Infant Dataset (MATINF) is the first large-scale dataset covering three major NLP tasks: text classification, question answering and summarization on Chinese.

LDC2004T07, LDC2005T06, LDC2005T10 and LDC2004T08 (Hongkong Hansards/Laws/News) to pre-train our model in sentence-level NMT. Then we employ the Ted talks corpus[4] to train our model on document-level NMT and use *dev2010* (8 documents with 879 sentence pairs) as the development corpus, *tst2012-2015* (62 documents with 5566 sentence pairs) as the test corpus.

For the QA and summarization tasks, we use the Maternal and Infant Dataset (Xu et al., 2020) for experimentation. The training corpus, validation corpus, and test corpus contain 0.75M, 0.21M, and 0.11M articles, respectively.

Model Settings. For document-level NMT, we apply our proposed approach to the Transformer model implemented by OpenNMT (Klein et al., 2017). For fair comparison, we keep our system settings the same as previous work (Tan et al., 2019), and the detailed model configurations are shown in Appendix. Following previous work, we also use the *multi-bleu.perl* script to compute case-insensitive BLEU score for evaluation.

For the QA and summarization tasks, we apply our proposed approach to the sequence-to-sequence model of MTF-S2S (the single task version) (Xu et al., 2020)[5], and keep the system settings consistent with (Xu et al., 2020). Concretely, we use the beam search algorithm during decoding, and we set the hidden size of encoders and decoders to 200 and the batch size to 64. We also use Adam as our optimizer with the learning rate set to 0.001. Similarly, we also use ROUGE (Lin and Hovy, 2003) to estimate the quality of the generated texts for performance evaluation.

4.2 Results on Document-level NMT

For document-level NMT, we compare our system with two recent context-aware systems (Zhang et al., 2018; Tan et al., 2019). Among them, Zhang et al. (2018) propose to model partial document context from previous sentences for better performance. And Tan et al. (2019) put their insight on global context modeling and have demonstrated the usefulness of global context in document-level NMT. Besides, we also present the results of Transformer (Vaswani et al., 2017) for comparison.

From the results in Table 1 we find that in comparison with the three baseline systems, our system (line 4) significantly outperforms (Tan et al., 2019) by 3.06 BLEU points and (Zhang et al., 2018) by 3.60 BLEU points on average. And the superiority is much significant when compared with the Transformer model. This indicates that coupling global context modeling with ZP recovering can significantly boost the document-level translation performance. Moreover, the ablation study (the last two lines) show that the adversarial method we use is definitely useful, although the performance improvement is not so significant.

Model	tst12	tst13	tst14	tst15
Transformer (2017)*	15.87	16.51	14.67	17.27
Zhang et al. (2018)*	16.46	17.80	15.85	18.24
Tan et al. (2019)*	16.94	18.31	16.21	19.07
Ours	20.06	21.22	19.00	22.47
- GAN	19.37	20.23	18.95	21.87

Table 1: Performance comparison on Chinese-English document-level NMT. The *p-value*s between our results and (Tan et al., 2019) are all less than 0.01. "*" denotes the results are borrowed from (Tan et al., 2019).

Model	R-1	R-2	R-L
Devlin et al. (2019)	18.66	3.28	10.78
Sutskever et al. (2014)	16.62	4.53	10.37
Luong et al. (2015)	19.62	5.87	13.34
Baseline (Xu et al., 2020)	20.28	5.94	13.52
Ours	22.16	6.26	14.61
- GAN	21.44	6.08	14.18

Table 2: Performance comparison on the QA task.

4.3 Results on QA & Summarization Tasks

For QA and summarization, we directly borrow the systems of Xu et al. (2020) as our baselines, and we perform experiments on their single version systems for clarity. Similar to (Xu et al., 2020), we also report the results of related systems on the two tasks for reference, and the results of these systems are directly borrowed from (Xu et al., 2020).

QA. For question answering, in addition to (Xu et al., 2020), we also compare with a retrieval-based baseline by fine-tuning BERT-base (Devlin et al., 2019) for question matching on an external dataset and two character-based generation baselines (Sutskever et al., 2014; Luong et al., 2015). The overall results are shown in Table 2. From the results, lines 4 and 5 show that our model outperforms the baseline (Xu et al., 2020) on all the three indicators. And the last two lines show that the

[4]Including 1906 documents with 226K sentence pairs from the IWSLT 2017 (Cettolo et al., 2012) evaluation campaigns https://wit3.fbk.eu.

[5]https://github.com/WHUIR/MATINF.

Model	R-1	R-2	R-L
Mihalcea and Tarau (2004)	35.53	25.78	36.84
Erkan and Radev (2004)	33.08	23.31	34.96
Sutskever et al. (2014)	23.05	11.44	19.55
Luong et al. (2015)	43.05	28.03	38.58
Ma et al. (2018)	34.63	22.56	28.92
Lin et al. (2018)	49.28	34.14	47.64
Liu and Lapata (2019)†	57.31	44.05	55.93
Baseline (Xu et al., 2020)	43.02	28.05	38.55
Ours	50.12	33.79	44.00
- GAN	49.82	33.34	43.44

Table 3: Performance comparison on summarization. † denotes the state-of-the-art BERT is used.

adversarial learning strategy we use still improves the QA performance to a certain extent. Notably, our resulting method has settled a new state-of-the-art performance on all the three indicators when compared with previous studies.

Summarization. For the summarization task, we compare with two extractive methods (Mihalcea and Tarau, 2004; Erkan and Radev, 2004) and six abstractive methods (Sutskever et al., 2014; Luong et al., 2015; Ma et al., 2018; Lin et al., 2018; Liu and Lapata, 2019; Xu et al., 2020). And the overall results are presented in Table 3. Firstly, compared with the baseline (Xu et al., 2020), our system significantly outperforms theirs by 16.50% on "R-1", 20.46% on "R-2", and 14.14% on "R-L", which proves the great effectiveness of coupling global context modeling with ZP recovering in text summarization. Secondly, compared with all previous studies, our system obtains a superior or competitive performance in most cases except BertAbs (Liu and Lapata, 2019) which employs a well-trained BERT model[6] for context-aware word representation. Similar to NMT and QA, the last two lines further demonstrate the usefulness of the adversarial learning strategy we utilize.

On the whole, the overall results above demonstrate that our proposed method is useful. Through this highly adaptive two-stage method, we can effectively alleviate ZP problems in various downstream NLG tasks with only a slight dependency on an independent small-scale corpus annotated with ZP positions. In addition, the results also show that mining the model's potential from a specific perspective (i.e., zero pronoun) is of great significance to document-level NLG. Naturally, we can also extent the proposed approach to other discourse phenomena (e.g., lexical cohesion, ellipsis, etc.) for discourse-aware language generation, which is worthy of in-depth study.

[6]We re-ran the TransformerAbs (Liu and Lapata, 2019) using the same word representation as ours but got a terrible result due to the differences in language and domain.

Model	APT	CRC
Transformer (2017)	54.34	45.18
Tan et al. (2019)	55.24	46.26
Ours	**71.92**	**56.13**

Table 4: Experimental results on pronoun translation. We evaluate on the previously mentioned four test sets and report the average score for comparison.

5 Analysis and Discussion

In this section, we aim at exploring the potential value of our proposed approach. For clarity, we analyze on document-level NMT for reference.

5.1 Contribution on Pronoun Translation

To comprehensively estimate the usefulness of our approach in pronoun translation, we employ two additional methods for evaluation, i.e., Accuracy of Pronoun Translation (APT) (Miculicich Werlen and Popescu-Belis, 2017) and Common Reference Context (CRC) (Jwalapuram et al., 2019).

On the one hand, we follow previous work (Tan et al., 2019) to use the APT[7] method to evaluate our pronoun generation performance. In addition, we also report the results of Transformer and Tan et al. (2019) for reference, as shown in Table 4. The results show that our model achieves a remarkable performance which significantly outperforms the Transformer by 32.4% and the context-aware model of Tan et al. (2019) by 30.20%. This strongly suggests the significant effectiveness of our proposed approach in pronoun translation. On the other hand, we further employ the novel CRC[8] method to evaluate the pronoun translation performance of our model, and the results are shown in Table 4. From the results we find that our proposed approach still significantly outperforms the Transformer model by 24.24% and the system of Tan et al. (2019) by 21.34%.

The overall results on both APT and CRC indicate that our method of coupling context modeling with ZP recovering is definitely effective and the

[7]A reference-based metric that measures the degree of overlapping pronouns between the output and reference translations obtained via word-alignments.

[8]An ELMo-based evaluation model is used to distinguish between good and bad translations via pair-wise ranking.

Model	tst12	tst13	tst14	tst15
Ctx→words	16.94	18.31	16.21	19.07
Ctx→pro	19.12	20.22	18.83	21.73
Ctx→pro&words	20.06	21.22	19.00	22.47

Table 5: Comparison of three ways of context information utilizing in document-level NMT.

Model	P	R	F1
Wang et al. (2016)*	82.48	74.38	78.22
Ours	84.04	76.31	79.99

Table 6: Performance of ZP position detection. "*": the re-produced performance on the data we use.

obtained ZP-focused context is expert in pronoun generation in document-level NLG tasks.

5.2 Different Strategies of Utilizing Global Document Context

As stated before, although Tan et al. (2019) have demonstrated the usefulness of global context, the context information is too complicated and it is hard to figure out what type of information is at work. To understand the role of context information, we explore the effects of different ways of using global context in document-level NMT. Concretely, we carry out experiments over three system settings where "Ctx→words" means distributing global context to word units (Tan et al., 2019), "Ctx→pro" means leveraging global context only for ZP recovery and then using the recovered ZP representation to replace the sentence-level hidden states of the placeholders, and "Ctx→pro&words" means leveraging global context for ZP recovery and distributing the ZP-focused context to word units. The overall results are shown in Table 5.

The first two lines in the table show that our recovered ZP representations (line 2) are effective due to the great capability of our approach in extracting more concise and effective features from the global context from a specific perspective (i.e., zero pronoun). Moreover, the last two lines show that distributing global context to word units can further improve the performance. This indicates that the global context still contains some other effective information worthy of further mining.

5.3 Performance of ZP Position Detection

In the literature, Wang et al. (2016) have achieved a certain success in ZP recovery in NMT. As stated before, we aim to improve their ZP position detector by considering both the left and right sides of each word as candidate ZP positions. To investigate the effect of our approach, we perform experiments on the cleaned `tvsub` corpus (Wang et al., 2018a) with the sentences without ZPs filtered out[9]. For performance evaluation, we follow Wang et al. (2016) to utilize the micro-averaged F_1-score to measure our model performance. The results in Table 6 show that the improved ZP position detector does achieve results better than the method of (Wang et al., 2018a), which suggests the necessity of taking both sides of each word into consideration for ZP position detection.

It is worth mentioning that since this work directly harnesses ZP representation in subsequent tasks aiming to alleviate error propagation, it does not depend on ZP-annotated NLG data, therefore, the evaluation on ZP label recovery is infeasible.

Type	Sentence
Source	[ni keyi dao wangzhan shang , xiazai suoyou de sheji wenjian , ziji lai zhizao **ZP-P** .]
Ref	[You can go on the website , download all the design files , make them yourself .]
Baseline	[You can go on the website and download all the design documents that make themselves .]
Ours	[You can go to the website , download all the design documents , make them yourself .]

Table 7: Comparison between example translation results of different methods. Here, "ZP-P" denotes an automatically predicted ZP position placeholder.

5.4 Case study

Here, we present a translation example of our NMT system in Table 7 for discussion. In the example, "Source" denotes a source sentence with ZP position detected; "Ref" denotes the reference translation; "Baseline" and "Ours" denote the translation results of the baseline system (Tan et al., 2019) and our approach, respectively. Referring to the "Ref" sentence, although the "Baseline" system can well leverage global context for better BLEU scores, the improvement on pronoun translation is still far from perfect. On the contrary, our system can accurately translate the pronoun "them" and the resulting sentence seems more fluent and more in line with the norms of the target language.

[9]We download `tvsub` from https://github.com/longyuewangdcu/tvsub, and the cleaned corpus have 389185 sentence pairs in total with 388346 pairs for training, 421 pairs for validation, and 418 pairs for testing.

6 Conclusion

In this paper, we introduced a highly adaptive two-stage method to mitigate the cohesion problem posed by the ZP phenomenon in document-level NLG tasks. To tackle both error propagation and corpus limitation issues, we first pre-trained a fault-tolerant ZP position detector for automatic ZP position prediction, and then performed document context modeling for both task-supervised ZP recovering and ZP-focused NLG task learning. And we also trained our model in an adversarial fashion to alleviate the language generation confusion caused by mis-recovered ZPs. Experiments on three D-NLG tasks show that our approach can greatly improve the performances, and the performance on pronoun translation is remarkable.

Acknowledgments

We would like to thank Professor Kong Fang for her valuable discussion on this work, and we also thank the anonymous reviewers for their insightful comments. This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600 and the National Natural Science Foundation of China (NSFC) via Grant Nos. 62076175 and 61876118.

References

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Susan Converse. 2006. Pronominal anaphora resolution in chinese. *Dissertations available from ProQuest*.

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 NAACL: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

G. Erkan and D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2964–2975, Hong Kong, China. Association for Computational Linguistics.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2242–2254, Online. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Fang Kong and Guodong Zhou. 2010. A tree kernel-based unified framework for Chinese zero anaphora resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 882–891, Cambridge, MA. Association for Computational Linguistics.

Charles Li and Sandra Thompson. 1979. *Third-Person Pronouns and Zero-Anaphora in Chinese Discourse*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the NAACL*, pages 150–157.

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. Global encoding for abstractive summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 163–169, Melbourne, Australia. 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 ACL*, pages 102–111, Vancouver, Canada. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 EMNLP*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the*

2015 EMNLP, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Shuming Ma, Xu Sun, Wei Li, Sujian Li, Wenjie Li, and Xuancheng Ren. 2018. Query and output: Generating words by querying distributed word representations for paraphrase generation. In *Proceedings of the 2018 NAACL: Human Language Technologies*, pages 196–206, New Orleans, Louisiana. Association for Computational Linguistics.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 NAACL: Human Language Technologies*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Sudha Rao, Allyson Ettinger, Hal Daumé III, and Philip Resnik. 2015. Dialogue focus tracking for zero pronoun resolution. In *Proceedings of the 2015 NAACL: Human Language Technologies*, pages 494–503, Denver, Colorado. Association for Computational Linguistics.

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 ACL*, pages 5429–5434, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Hirotoshi Taira, Katsuhito Sudoh, and Masaaki Nagata. 2012. Zero pronoun resolution can improve the quality of j-e translation. In *Proceedings of the 6th Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 111–118, Jeju, Republic of Korea. Association for Computational Linguistics.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 1576–1585, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018a. Translating pro-drop languages with reconstruction models. pages 4937–4945.

Longyue Wang, Zhaopeng Tu, Xing Wang, and Shuming Shi. 2019. One model to learn both: Zero pronoun prediction and translation. In *Proceedings of the 2019 EMNLP*, pages 921–930, Hong Kong, China. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2018b. 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 NAACL: Human Language Technologies*, pages 983–993, San Diego, California. Association for Computational Linguistics.

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. 2020. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 5971–5978, Online. Association for Computational Linguistics.

Bing Xiang, Xiaoqiang Luo, and Bowen Zhou. 2013. Enlisting the ghost: Modeling empty categories for machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 822–831, Sofia, Bulgaria. Association for Computational Linguistics.

Canwen Xu, Jiaxin Pei, Hongtao Wu, Yiyu Liu, and Chenliang Li. 2020. MATINF: A jointly labeled large-scale dataset for classification, question answering and summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3586–3596, Online. Association for Computational Linguistics.

Jingxuan Yang, Kerui Xu, Jun Xu, Si Li, Sheng Gao, Jun Guo, Ji-Rong Wen, and Nianwen Xue. 2020. Transformer-GCRF: Recovering Chinese dropped pronouns with general conditional random fields. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 137–147, Online. Association for Computational Linguistics.

Yaqin Yang, Yalin Liu, and Nianwen Xue. 2015. Recovering dropped pronouns from Chinese text messages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 309–313, Beijing, China. Association for Computational Linguistics.

Qingyu Yin, Yu Zhang, Wei-Nan Zhang, Ting Liu, and William Yang Wang. 2018. Deep reinforcement learning for Chinese zero pronoun resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 569–578, Melbourne, Australia. Association for Computational Linguistics.

Hongming Zhang, Yan Song, and Yangqiu Song. 2019a. Incorporating context and external knowledge for pronoun coreference resolution. In *Proceedings of the 2019 NAACL: Human Language Technologies*, pages 872–881, Minneapolis, Minnesota. Association for Computational Linguistics.

Hongming Zhang, Yan Song, Yangqiu Song, and Dong Yu. 2019b. Knowledge-aware pronoun coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 867–876, Florence, Italy. Association for Computational Linguistics.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Weinan Zhang, Ting Liu, Qingyu Yin, and Yu Zhang. 2019c. Neural recovery machine for chinese dropped pronoun. *Frontiers of Computer Science*, 13(5):1023–1033.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of Chinese zero pronouns: A machine learning approach. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing*, pages 541–550, Prague, Czech Republic. Association for Computational Linguistics.

Appendices

A. Model Configuration

For the NMT models, following the parameter settings of Transformer, we set the hidden size and filter size to 512 and 2048 respectively. We set the layer number of encoder and decoder to 6 and used 8 attention heads in the multi-head attention of each layer. We set the source and target vocabulary size as 50K and each batch contains 4096 tokens. We set the beam size and dropout rate to 5 and 0.1 respectively, and the settings on the Adam optimization and regularization methods were the same as Transformer. Notably, we manually set the above parameter values as previous work only for fair comparison. We trained the model for a total of 125,000 steps on the GeForce RTX 2080Ti GPUs and each training step took about 1.5 seconds, and the number of parameters in our model was around 120M.

B. Pronoun Labels

We take the 30 kinds of Chinese pronoun categories and the "ε" category into consideration in our experiments, as presented in Table 8.

我 (*I, me*), 我们 (*we, us*), 你/你们 (*you, you*), 他 (*he, him*), 她 (*she, her*), 它 (*it, it*), 他们/她们/它们 (*they, them*)
我的 (*my, mine*), 我们的 (*our, ours*), 你的/你们的 (*your, yours*), 他的 (*his, his*), 她的 (*her, hers*), 它的 (*its, its*), 他们的/她们的/它们的 (*their, theirs*)
我自己 (*myself*), 我们自己 (*ourselves*), 你自己 (*yourself*), 你们自己 (*yourselves*), 他自己 (*himself*), 她自己 (*heirself*), 它自己 (*itself*), 他们自己/她们自己/他们自己 (*themselves*)
空 (ε)

Table 8: Pronoun labels used in our experiments.

C. Effects of Document-Level Encoder with Different Layer Numbers

In this study, the extracted document context plays an important role in both ZP recovering and NLG task learning. Therefore, we conduct experiments on the document-level encoder over different layer numbers, and the number of search trials is around 8. As shown in Table 9, the layer number of the document-level encoder does not make much difference in our system. According to the results, we set the layer number to 6 in our experiments.

layer	tst12	tst13	tst14	tst15	avg
6	20.06	21.22	19.00	22.47	20.69
5	19.69	20.43	19.02	21.90	20.26
4	19.95	20.63	18.93	22.06	20.39
3	20.14	21.03	18.96	21.95	20.52
2	19.78	20.92	19.44	22.44	20.65
1	20.21	20.93	19.33	22.24	20.68

Table 9: NMT results over different layer numbers.