One Model to Learn Both: Zero Pronoun Prediction and Translation

Longyue Wang
Tencent AI Lab
vinnylywang@tencent.com

Zhaopeng Tu
Tencent AI Lab
zptu@tencent.com

Xing Wang
Tencent AI Lab
brightxwang@tencent.com

Shuming Shi
Tencent AI Lab
shumingshi@tencent.com

Abstract

Zero pronouns (ZPs) are frequently omitted in pro-drop languages, but should be recalled in non-pro-drop languages. This discourse phenomenon poses a significant challenge for machine translation (MT) when translating texts from pro-drop to non-pro-drop languages. In this paper, we propose a unified and discourse-aware ZP translation approach for neural MT models. Specifically, we jointly learn to predict and translate ZPs in an end-to-end manner, allowing both components to interact with each other. In addition, we employ hierarchical neural networks to exploit discourse-level context, which is beneficial for ZP prediction and thus translation. Experimental results on both Chinese⇒English and Japanese⇒English data show that our approach significantly and accumulatively improves both translation performance and ZP prediction accuracy over not only baseline but also previous works using external ZP prediction models. Extensive analyses confirm that the performance improvement comes from the alleviation of different kinds of errors especially caused by subjective ZPs.

1 Introduction

Zero anaphora is a discourse phenomenon, where pronouns can be omitted when they are pragmatically or grammatically inferable from intra- and inter-sentential context (Li and Thomson, 1979). However, translating such implicit information (i.e. zero pronoun, ZP) poses various difficulties for machine translation (MT) in terms of completeness and correctness. Although neural models are getting better at learning representations, it is still difficult to implicitly learn complex ZPs in a general model. Actually, ZP prediction and translation need to not only understand the semantics or intentions of a single sentence, but also utilize its discourse-level context.

Two technological advances in the field of ZP and MT, have seen vast progress over the last decades, but they have been developed very much in isolation. Early studies (Chung and Gildea, 2010; Le Nagard and Koehn, 2010; Xiang et al., 2013) fed MT systems with the results of ZP prediction models, which are trained on a small-scale and non-homologous data compared to MT models. To narrow the data-level gap, Wang et al. (2016) proposed an automatic method to annotate ZPs by utilizing the parallel corpus of MT. The homologous data for both ZP prediction and translation leads to significant improvements on translation performances for both statistical MT (Wang et al., 2016) and neural MT models (Wang et al., 2018a). However, such approaches still require external ZP prediction models, which have a low accuracy of 66%. The numerous errors of ZP prediction errors will be propagated to translation models, which leads to new translation problems. In addition, relying on external ZP prediction models in decoding makes these approaches unwieldy in practice, due to introducing more computation cost and pipeline complexity.

In this work, we try to further bridge the model-level gap by jointly modeling ZP prediction and translation. Joint learning has proven highly effective on alleviating the error propagation problem, such as joint parsing and translation (Liu and Liu, 2010), as well as joint tokenization and translation (Xiao et al., 2010). Similarly, we expect that ZP prediction and translation could interact with each other: prediction offers more ZP information beyond 1-best result to translation and translation helps prediction resolve ambiguity. Specifically, we first cast ZP prediction as a sequence labeling task with a neural model, which is trained jointly with a standard neural machine translation (NMT) model in an end-to-end manner. We leverage the auto-annotated ZPs to supervise the learning of ZP

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 921–930,
Hong Kong, China, November 3–7, 2019.

prediction component, which releases the reliance on external ZP knowledge in decoding phase.

In addition, previous studies revealed that discourse-level information can better tackle ZP resolution, because around 23% of ZPs appear two or more sentences away from their antecedents (Zhao and Ng, 2007; Chen and Ng, 2013). Inspired by these findings, we exploit inter-sentential context to further improve ZP prediction and thus translation. Concretely, we employ hierarchical neural networks (Sordoni et al., 2015; Wang et al., 2017) to summarize the context of previous sentences in a text, which is integrated to the joint model for ZP prediction.

We validate the proposed approach on the widely-used data for ZP translation (Wang et al., 2018a), which consist of 2.15M Chinese–English sentence pairs. Experimental results show that the joint model indeed improves performances on both ZP prediction and translation. Incorporating discourse-level context further improves performances, and outperforms ther external ZP prediction model (Wang et al., 2018a) by +2.29 BLEU points in translation and +11% in prediction accuracy. Experimental results on a further Japanese–English translation task show that our model consistently outperforms both the baseline and the external ZP prediction model, demonstrating the universality of the proposed approach.

The key contributions of this paper are:

1. We propose a single model to jointly learn ZP prediction and translation, which improves performances on both tasks by allowing the two components to interact with each other.

2. Our study demonstrates the effectiveness of discourse-level context for ZP prediction.

3. Based on our manually-annotated testset, we conduct extensive analyses to assess ZP prediction and translation.

2 Background

2.1 Zero Pronoun

In pro-drop languages such as Chinese and Japanese, ZPs occur much more frequently compared to non-pro-drop languages such as English (Zhao and Ng, 2007). As seen in Table 1, the subject pronoun ("我") and the object pronoun ("它") are omitted in Chinese sentences ("Inp.") while these pronouns are all compulsory in their English translations ("Ref."). This is not a problem for human beings since we can easily recall these missing pronoun from the context. Taking the second sentence for example, the pronoun "它" is an anaphoric ZP that refers to the antecedent ("蛋糕") in previous sentence, while the non-anaphoric pronoun "我" can still be inferred from the whole sentence. The first example also indicates the necessity of intra-sentential information for ZP prediction.

Inp.	等 我 搬进来，**(我)** 能 买 台 电视 吗?
Ref.	Can **I** get a TV when I move in?
Out.	When I move in **to buy** a TV.
Inp.	这块 蛋糕 很 美味！ 你 烤 的 **(它)** 吗?
Ref.	The cake is very tasty! Did you bake **it**?
Out.	The cake is delicious! **Are you baked**?

Table 1: Examples of ZPs and translations where words in brackets are ZPs that are invisible in decoding and underlined words are antecedents of anaphoric ZPs. This leads to problems for NMT in respect of completeness (first case) and correctness (second case). "Inp." and "Ref." indicate Chinese input and English translation, respectively. "Out." represents the output of a NMT model.

However, ZP poses a significant challenge for translation models from pro-drop to non-pro-drop languages, where ZPs are normally omitted in the source side but should be generated overly in the target side. As shown in Table 1, even a strong NMT model fails to recall the implicit information, which lead to problems like *incompleteness* and *incorrectness*. The first case is translated into "When I move in to buy a TV", which makes the output miss subject element (incompleteness). The second case is translated into "Are you baked?", while the correct translation should be "Did you bake it?" (incorrectness).

2.2 Bridging Data Gap Between ZP Prediction and Translation

Recent efforts have explored ways to bridge the gap of ZP prediction and translation (Wang et al., 2016, 2018a,b) by training both models on the homologous data. The pipeline involves two phases, as described below.

Translation-Oriented ZP Prediction Its goal is to recall the ZPs in the source sentence (i.e. pro-drop language) with the information of the target sentence (i.e. non-pro-drop language) in a paral-

lel corpus. Taking the second case (assuming that Inp. and Ref. are sentence pair in a parallel corpus) in Table 1 for instance, the ZP "它 (*it*)" is dropped in the Chinese side while its equivalent "it" exists in the English side. It is possible to identify the ZP position (between "的" and "吗") by alignment information, and then recover the ZP word "它" by a language model (scoring all possible pronoun candidates and select the one with the lowest perplexity). Wang et al. (2016) proposed a novel approach to automatically annotate ZPs using alignment information from bilingual data, and the auto-annotation accuracy can achieve above 90%. Thus, a large amount of ZP-annotated sentences were available to train an external ZP prediction model, which was further used to annotate source sentences in test sets during the decoding phase. They integrated the ZP predictor into SMT and showed promising results on both Chinese–English and Japanese–English data.

However, their neural-based ZP prediction model still produce low accuracies on predicting ZPs, which is 66% in F1 score. This is a key problem for the pipeline framework, since numerous errors would be propagated to the subsequent translation process.

Translation with ZP-Annotated Data An intuitive way to exploit the annotated data is to train a standard NMT model on the annotated parallel corpus, which decodes the input sentence annotated by the external ZP prediction model. Wang et al. (2018a) leveraged the encoder-decoder-reconstructor framework (Tu et al., 2017) for this task, which reconstructs the intermediate representations of NMT model back to the ZP-annotated input. The auxiliary loss on ZP reconstruction can guide the intermediate representations to learn critical information relevant to ZPs. However, their best model still needs external ZP prediction at decoding time. In response to this problem, Wang et al. (2018b) leveraged the prediction results of the ZP positions, which have relatively higher accuracy (e.g. 88%). Accordingly, they jointly learn the partial ZP prediction (*i.e.*, predict the ZP word given the externally annotated ZP position) and ZP translation.

In this work, we follow this direction with the encoder-decoder-reconstructor framework, and show our approach outperforms both strategies of using externally annotated data.

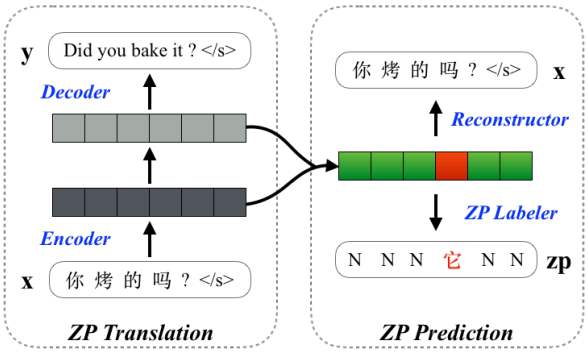

Figure 1: Architecture of the joint ZP prediction and translation model, in which ZP prediction is casted as a sequence labelling problem.

3 Approach

In this study, we propose a joint model to learn ZP prediction and translation, which can be further improved by leveraging discourse-level context.

- *Joint ZP Prediction and Translation* (Section 3.1) We cast ZP prediction as a sequence labeling problem, which can be trained together with ZP translation model in an end-to-end manner. This releases the reliance on external ZP prediction models (e.g. 66% or 88% accuracy), since no ZP-annotated sentence is required any more in decoding. Instead, only the high-quality annotated bilingual data (e.g. 93% accuracy) are needed.

- *Discourse-Aware ZP Prediction* (Section 3.2) We further improve ZP prediction and thus its translation with discourse-level context, which is summarized by hierarchical neural networks. The contextual representation is integrated into the reconstructor, based on which ZP prediction is conducted.

3.1 Joint ZP Prediction and Translation

Figure 1 illustrates the architecture of the joint model, which consists of two main components. The ZP translation component is a standard encoder-decoder NMT model, while an additional reconstructor is introduced for ZP prediction. To guarantee the reconstructor states contain enough information for ZP prediction, the reconstructor reads both the encoder and decoder states and the reconstruction score is computed by

$$R(\hat{\mathbf{x}}|\mathbf{h}^{enc}, \mathbf{h}^{dec}) = \prod_{t=1}^{T} g_r(\hat{x}_{t-1}, \mathbf{h}_t^{rec}, \hat{\mathbf{c}}_t^{enc}, \hat{\mathbf{c}}_t^{dec})$$

where $\mathbf{h}_t^{rec}$ is the hidden state in the reconstructor:

$$\mathbf{h}_t^{rec} = f_r(\hat{x}_{t-1}, \mathbf{h}_{t-1}^{rec}, \hat{\mathbf{c}}_t^{enc}, \hat{\mathbf{c}}_t^{dec}) \quad (1)$$

Here $g_r(\cdot)$ and $f_r(\cdot)$ are respective softmax and activation functions for the reconstructor. The context vectors $\hat{\mathbf{c}}_t^{enc}$ and $\hat{\mathbf{c}}_t^{dec}$ are the weighted sum of $\mathbf{h}^{enc}$ and $\mathbf{h}^{dec}$, and the weights are calculated by two interactive attention models:

$$\hat{\alpha}^{enc} = \text{ATT}_{enc}(x_{t-1}, \mathbf{h}_{t-1}^{rec}, \mathbf{h}^{enc}) \quad (2)$$
$$\hat{\alpha}^{dec} = \text{ATT}_{dec}(x_{t-1}, \mathbf{h}_{t-1}^{rec}, \mathbf{h}^{dec}, \hat{\mathbf{c}}_t^{enc}) \quad (3)$$

The interaction between two attention models leads to a better exploitation of the encoder and decoder representations (Wang et al., 2018b).

ZP Prediction as Sequence Labelling We cast ZP prediction as a sequence labelling task, where each word is labelled if there is a pronoun missing before it. Given the input $\mathbf{x} = \{x_1, x_2, \ldots, x_T\}$ with the last word x_T being the end-of-sentence tag " ⟨eos⟩",[1] the output to be labelled is a sequence of labels $\mathbf{zp} = \{zp_1, zp_2, \ldots, zp_T\}$ with $zp_t \in \{N\} \cup \mathbb{V}_{zp}$. Among the label set, "$N$" denotes no ZP, and $\mathbb{V}_{zp}$ is the vocabulary of pronouns.[2] Taking Figure 1 as an example, the label sequence "N N N 它 N N" indicates that the pronoun "它" is missing before the fourth word "吗" in the source sentence "你 烤 的 吗？ ". More specifically, we model the probability of generating the label sequence $\mathbf{zp}$ as:

$$P(\mathbf{zp}|\mathbf{h}^{rec}) = \prod_{t=1}^{T} P(zp_t|\mathbf{h}_t^{rec}) = \prod_{t=1}^{T} g_l(zp_t, \mathbf{h}_t^{rec}) \quad (4)$$

where $g_l(\cdot)$ is softmax for the ZP labeler. As seen, we integrate the ZP generation component into the ZP translation model. There is no reliance on external ZP prediction models in decoding phase.

Training and Testing The newly introduced prediction component is trained together with the encoder-decoder-reconstructor:

$$J(\theta, \gamma, \psi) = \arg\max_{\theta,\gamma,\psi} \Big\{ \underbrace{\log L(\mathbf{y}|\mathbf{x};\theta)}_{\textit{likelihood}} + \underbrace{\log R(\mathbf{x}|\mathbf{h}^{enc}, \mathbf{h}^{dec};\theta)}_{\textit{reconstruction}} + \underbrace{\log P(\mathbf{zp}|\mathbf{h}^{rec};\theta,\gamma)}_{\textit{ZP labeling}} \Big\} \quad (5)$$

where $\{\theta, \gamma\}$ are respectively the parameters associated with the encoder-decoder-reconstructor and the ZP prediction component. The auxiliary prediction loss $P(\cdot)$ guides the hidden states of both the encoder-decoder and the reconstructor to embed the ZPs in the source sentence. Although the calculation of labeling loss relies on explicitly annotated labels, it is only used in training to guide the parameters to learn ZP-enhanced representations. Benefiting from the implicit integration of ZP information, we release the reliance on external ZP prediction model in testing.

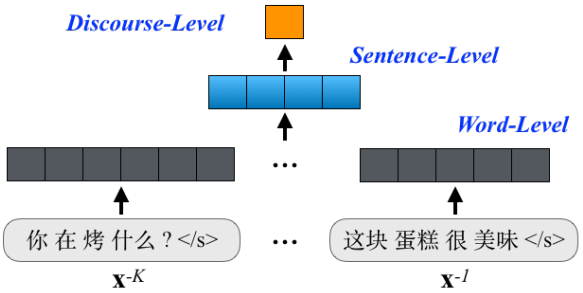

Figure 2: Architecture of hierarchical neural encoder. $\mathbf{x}^{-K}, \ldots, \mathbf{x}^{-1}$ are K previous sentences before the current source sentence "你 烤 的 吗 ?" in a text.

3.2 Discourse-Aware ZP Prediction

Discourse information have proven useful for predicting antecedents, which may occur in previous sentences (Zhao and Ng, 2007; Chen and Ng, 2013). Therefore, we further improve ZP prediction with discourse-level context, which is learned together with the joint model.

Encoding Discourse-Level Context Hierarchical structure networks are usually used for modelling discourse context on various natural language processing tasks such query suggestion (Sordoni et al., 2015), dialogue modeling (Serban et al., 2016) and MT (Wang et al., 2017). Therefore, we employ hierarchical encoder (Wang et al., 2017) to encoder discourse-

[1] We introduce " ⟨eos⟩" to cover the case that a pronoun is missing at the end of a sentence.

[2] We employ the pronoun vocabulary used in Wang et al. (2016), which contains 30 distinct Chinese pronouns.

level context for NMT. More specifically, we use the previous K source sentences $\mathbf{X} = \{\mathbf{x}^{-K}, \ldots, \mathbf{x}^{-1}\}$ as the discourse information, which is summarized with a two-layer hierarchical encoder, as shown in Figure 2. For each sentence $\mathbf{x}^{-k}$, we employ a *word-level encoder* to summarize the representation of the whole sentence:

$$\mathbf{h}^{-k} = \text{ENCODER}_{word}(\mathbf{x}^{-k}) \quad (6)$$

After we can obtain all sentence-level representations $\mathbf{H}^X = \{\mathbf{h}^{-K}, \ldots, \mathbf{h}^{-1}\}$, we feed them into a *sentence-level encoder* to produce a vector that represents the discourse-level context:

$$\mathbf{C} = \text{ENCODER}_{sentence}(\mathbf{H}^X) \quad (7)$$

Here the summary C consists of not only the dependencies between words, but also the relations between sentences. Following Voita et al. (2018), we share the parameters of word-level encoder ENCODER_{word} with the encoder component in standard NMT model. Note that, ENCODER_{word} and $\text{ENCODER}_{sentence}$ can be implemented as arbitrary networks, such as recurrent networks (Cho et al., 2014), convolutional networks (Gehring et al., 2017), or self-attention networks (Vaswani et al., 2017). In this study, we used recurrent networks to implement our ENCODER.

Integrating Discourse into ZP Prediction We directly feed the discourse-level context to the reconstructor to improve ZP prediction. Specifically, we combine the context vector and the reconstructor state:

$$\widehat{\mathbf{h}}_t^{rec} = f_c(\mathbf{h}_t^{rec}, \mathbf{C}) \quad (8)$$

Here $f_c(\cdot)$ is a function for combining reconstructor states and the context vector, which is a simple concatenation (CONCAT) in this work. The revised reconstructor state $\widehat{\mathbf{h}}_t^{rec}$ is then used in Equations (1) and (4).

4 Experiments

4.1 Setup

We conducted translation experiments on both Chinese⇒English and Japanese⇒English translation tasks, since Chinese and Japanese are pro-drop languages while English is not. For Chinese⇒English translation task, we used the data of auto-annotated ZPs (Wang et al., 2018a).[3] The training, validation, and test sets contain 2.15M, 1.09K, and 1.15K sentence pairs, respectively. In the training data, there are 27% of Chinese pronouns are ZPs, which poses difficulties for NMT models. For Japanese⇒English translation task, we respectively selected 1.03M, 1.02K, and 1.02K sentence pairs from Opensubtitle2016[4] as training, validation, and test sets (Tiedemann, 2012). We used case-insensitive 4-gram NIST BLEU (Papineni et al., 2002) as evaluation metrics, and *sign-test* (Collins et al., 2005) to test for statistical significance.

To make fair comparison with Wang et al. (2018a), we also implemented our approach on top of the RNN-based NMT model, which incorporates dropout (Hinton et al., 2012) on the output layer and improves the attention model by feeding the most recently generated word. For training the models, we limited the source and target vocabularies to the most frequent 30K words for Chinese⇒English and 20K for Japanese⇒English. Each model was trained on sentences of length up to a maximum of 20 words with early stopping. Mini-batches were shuffled during processing with a mini-batch size of 80. The dimension of word embedding was 620 and the hidden layer size was 1,000. We trained for 20 epochs using Adadelta (Zeiler, 2012), and selected the model that yielded best performances on validation sets. For training the proposed models, the hidden layer sizes of hierarchical model and reconstruction model are 1,000 and 2,000, respectively. We modeled previous three sentences as discourse-level context.[5]

4.2 Results on Chinese⇒English Task

Table 2 lists the performance of ZP translation and prediction on Chinese⇒English data.

The baseline (Row 1) is trained on the standard NMT model using the original parallel data ($\mathbf{x}$, $\mathbf{y}$). In addition, we implemented two comparative models (Row 2-3), which differ with respect to the training data used. The "+ ZP-Annotated Data" model was still trained on standard NMT model but using new training instances ($\hat{\mathbf{x}}$, $\mathbf{y}$) whose source-side sentences are auto-annotated with ZPs. The "+ Reconstruction" is the best model reported in Wang et al. (2018a), which employs two reconstructors to reconstruct the $\hat{\mathbf{x}}$ from

[3] https://github.com/longyuewangdcu/tvsub.

[4] http://www.opensubtitles.org.

[5] We followed Wang et al. (2017) and Tu et al. (2018) to use 3 previous sentences as discourse context.

#	Model	Translation		Prediction		
		#Params	*BLEU*	*P*	*R*	*F1*
1	Baseline	86.7M	31.80	n/a	n/a	n/a
External ZP Prediction (Wang et al., 2018a)						
2	+ ZP-Annotated Data	+0M	32.67	0.67	0.65	0.66
3	+ Reconstruction	+73.8M	35.08			
This Work: Joint ZP Prediction and Translation						
4	Joint Model	+35.6M	36.04†	0.72	0.68	0.70
5	+ Discourse-Level Context	+56.6M	**37.11**†	0.76	0.77	**0.77**

Table 2: Evaluation of ZP translation and prediction on the Chinese–English data. "#Params" represents the number of parameters used in different models. "†" indicates statistically significant difference ($p < 0.01$) from the best external ZP prediction model for translation performance. As seen, the proposed joint models improve performances in both ZP translation and prediction, over the external ZP prediction models.

hidden representations of encoder and decoder. At decoding time, ZPs can not be annotated by alignment method since target sentences are not available. Thus, source sentences are annotated by an external ZP prediction model, which is trained on monolingual training instances $\hat{\mathbf{x}}$. Finally, we evaluated two proposed models (Row 4-5) which are introduced in Section 3.1 and 3.2, respectively.

Translation Quality Benefiting from the explicitly annotated ZPs in the source language, the "+ ZP-Annotated Data" model (Row 2) outperforms the baseline system built on the original data where the pronouns are missing (*i.e.,* +0.87 BLEU point). This illustrates that explicitly recalling translation of ZPs at training time helps produce better translations. Furthermore, the "+ Reconstuction" approach (Row 3) respectively outperforms the baseline and "+ ZP-Annotated Data" models by +3.28 and +2.41 BLEU points, which indicates that explicitly handling ZPs with reconstruction model can better address ZP problems.

The proposed models consistently outperform other models in all cases, demonstrating the superiority of the joint learning of ZP prediction and translation. Specifically, the "Joint Model" (Row 4) significantly improves translation performance by +4.24 over baseline model. In addition, this joint approach also outperforms two comparative models "+ ZP-Annotated Data" and "+ Reconstruction" by +3.37 and +0.96 BLEU points, respectively. We attribute the improvement over external ZP prediction to: 1) releasing the reliance on external ZP prediction models can greatly alleviate error propagation problems; and 2) joint learning of ZP prediction and translation is able to guide the related parameters to learn better latent representations. Furthermore, introducing discourse-level context (Row 5) accumulatively improves translation performance, and significantly outperform the joint model by +1.07 BLEU points.

Model	BLEU	△
Baseline	19.94	–
External ZP Prediction	20.86	+0.92
Joint Model	21.39	+1.45
+ Discourse-Level Context	**22.00**	+2.06

Table 3: Translation quality on Japanese–English data. As seen, the proposed models can also significantly improve translation performance, which shares the same trend with that on Chinese–English translation.

More parameters may capture more information, at the cost of posing difficulties to training. Wang et al. (2018a) leverage two separate reconstructors with hidden state size being 2000 and 1000 respectively. Accordingly, their models introduce a large number of parameters. In contrast, we set the hidden size of the reconstructor be 1000, which greatly reduce the newly introduced parameters (+35.6M vs. +73.8M). Modeling discourse-level context further introduces +21M new parameters, which is reasonable comparing with previous work. Our best model variation outperform that of external ZP prediction by over 2 BLEU points with less parameters (143.3M vs. 160.5M), showing that the improvements are attributed to the stronger modeling capacity rather than more parameters.

ZP Prediction Accuracy The joint model improves prediction accuracy as expected, which we

Model	ZP-Annotated Input		
	✓	×	▽
Baseline	31.80		–
External ZP Predict.	35.08	34.02	-1.06
Joint Model	36.04	35.93	-0.11
+ Discourse	**37.11**	**36.51**	-0.60

Table 4: Translation results when no ZP-annotated input is used in decoding by *removing the reconstructor component*. "▽" denotes the performance gap between whether using the annotated input ("✓") or not ("×").

Model	BLEU	△
Baseline	31.80	–
+ Discourse⇒Decoder	32.34	+0.54
Baseline + ZP-Anno.	32.67	
+ Discourse⇒Decoder	32.55	-0.12
Joint Model	36.04	–
+ Discourse⇒Decoder	34.66	-1.38

Table 5: Translation results when transforming the contextual representation to decoder of different models. Incorporating discourse-level context does not always lead to improvement of translation performance.

attribute to the leverage of useful translation information. Incorporating the discourse-level context further improves ZP prediction, and the best performance is 11% higher than external ZP prediction model. These results confirm our claim that joint learning of ZP prediction and translation can benefit both components by allowing them to interact with each other.

4.3 Results on Japanese⇒English Task

Table 3 lists the results. We compare our models and the best external ZP prediction approach. As seen, our models also significantly improve translation performance, demonstrating the effectiveness and universality of the proposed approach.

This improvement on Japanese⇒English translation is lower than that on Chinese⇒English, showing that ZP prediction and translation are more challenging for Japanese. The reason may be two folds: 1) Japanese language has a larger number of pronoun variations borrowed from archaism, which leads to more difficulties in learning ZPs; 2) Japanese language is subject-object-verb (SOV) while English has subject-verb-object (SVO) structure, and this poses difficulties for ZP annotation via alignment method.

4.4 Analysis

We conducted extensive analyses on Chinese ⇒English to better understand our models in terms of the effect of external ZP annotation and different types of ZPs errors.

Reliance on Externally ZP-Annotated Input Some researchers may argue that previous approaches (Wang et al., 2018a) are also able to release the reliance of externally annotated input by removing the reconstructor component. Table 4 lists the results. Without ZP-annotated input in decoding, all approaches can still outperform the baseline model, by benefiting better intermediate representations that contain necessary ZP information. Compared with reconstruction-based models, however, removing the reconstruction components leads to decrease on translation quality. As seen, the BLEU score of best "External ZP prediction" model dramatically drops by -1.06 points, showing that this approach is heavily dependent on the results of external ZP annotations. The performances of proposed models only decrease by -0.1∼-0.6 BLEU point. It indicates that our models are compatible with the standard encoder-decoder-reconstructor framework, thus enjoy an additional benefit of re-scoring translation hypotheses in testing with reconstruction scores. All the results together prove the superiority of the proposed unified framework for ZP translation.

Effect of Discourse-Level Context Recent studies revealed that inter-sentential context can implicitly help to tackle anaphora resolution in NMT architecture (Jean et al., 2017b; Bawden et al., 2018; Voita et al., 2018). Some may argue that document-level architectures are strong enough to alleviate ZP problems for NMT. To answer this concern, we compared with "+ Discourse⇒Decoder" models, which transform the contextual representation to the decoder part of different models. In this way, the discourse-level context can benefit both the generation of translation and ZP prediction.

As shown in Table 5, directly incorporating inter-sentential context into standard NMT model (one of document-level NMT architectures) can improve translation quality by +0.54 BLEU point than baseline. However, this integration mechanism does not work well in "Baseline + ZP-Annotation" and our "Joint" models, which de-

Model	Error	Sub.	Obj.	Dum.	All
BASE.	Total	112	41	45	198
EXTE.	Fixed	50	34	33	117
	New	11	14	7	32
	Total	73	21	19	113
JOIN.	Fixed	61	35	37	133
	New	8	11	7	26
	Total	59	17	15	91
+DIS.	Fixed	70	39	38	147
	New	7	9	7	23
	Total	**49**	**11**	**14**	**74**

Table 6: Translation error statistics. The ZP types "Sub.", "Obj." and "Dum." denote errors caused by subjective, objective and dummy pronouns, respectively. The models "Base.", "Exte.", "Join." and "+Dis." denote "Baseline", "+ Reconstruction", "Joint Model" and "+ Discourse-Level context" models. **Bold** numbers denote the least errors in each category.

creasing by -0.12 and -1.38 BLEU points, respectively. One potential problem with this strategy is that the propagation path is longer: $\mathbf{C} \rightarrow \mathbf{h}^{dec} \rightarrow \mathbf{h}^{rec} \rightarrow \mathbf{zp}$, which may suffer from the vanishing effect. This also confirms our hypothesis that discourse-level context benefits ZP prediction more than ZP translation. Therefore, we incorporate the discourse-level context into reconstructor instead of the decoder.

Manual Evaluation on Translation Errors We finally investigate how the proposed approaches improve the translation by human evaluation. We randomly select 500 sentences from the test set. As shown in Table 6, we count how many translation errors caused by different types of ZPs (*i.e.,* "Subjective", "Objective" and "Dummy"[6]) are fixed ("Fixed") and newly generated ("New") by different models.

All the models can fix different amount of ZP problems in terms of completeness and correctness, which is consistent with the translation results reported in Table 2. This confirms that our improvement in terms of BLEU scores indeed comes from alleviating translation errors caused by ZPs. Among them, the proposed model "+DIS." performs best, which fixes 74% of the ZP errors, and only introduces 12% of new errors.

In addition, we found that subjective ZPs are more difficult to predict and translate since they usually occur in imperative sentences, and ZP prediction needs to understand intention of speakers. The "EXTE." model only fixes 45% of subjective ZP errors but made 10% new errors by predicting wrong ZPs. However, the proposed joint model works better, which fixes 54% error with only introducing 7% new errors. Predicting objective ZPs needs inter-sentential context, thus our "+DIS." model is able to fix more objective ZP errors (95% vs. 82%) by introducing less new errors (22% vs. 34%) than "EXTE.".

Case Study Table 7 shows two typical examples, of which pronouns are mistakenly translated by the strong baseline ("External ZP Prediction") model (Wang et al., 2018a) while fixed by our model and failed to be fix. In "Fixed Error" case, the dropped word "它 (*it*)" is an anaphoric ZP whose antecedent is the noun "电视 (*television*)" in previous sentence while the dropped word "你 (*you*)" is a non-anaphoric ZP that depends upon speaker or listener. As seen, our "JOIN." model performs better than the "EXTE." model because two ZP positions are syntactically recalled in the target side, showing that the joint approach have better capability of utilizing intra-sentential information for identifying ZPs. Besides, our "+DIS." model can semantically fix the error by predicting correct ZP words, demonstrating that inter-sentential context can aid to recovering such complex ZPs. However, as shown in "Non-Fixed Error" case, there are still some ZPs can not be precisely predicted due to the misunderstanding of intentions of utterances. Thus, exploiting dialogue focus for ZP translation is our future work (Rao et al., 2015).

5 Related Work

ZP Prediction and Translation ZP resolution is a challenging task which needs lexical, syntactic, discourse knowledge. Previous studies have been conducted to improves the performance of ZP resolution for different pro-drop languages (Kong and Zhou, 2010; Chen and Ng, 2013; Park et al., 2015; Yin et al., 2017). However, directly using results of external ZP resolution systems for translation task shows limited improvements (Chung and Gildea, 2010; Le Nagard and Koehn, 2010; Taira et al., 2012; Xiang et al., 2013), since such external systems are trained on small-scale data that is non-homologous to MT. To

[6]In pro-drop languages, it is used to fulfill the syntactical requirements without providing explicit meaning (e.g. "it").

Fixed Error	
PRE.	等 我 搬进 来, 能 买 台 电视 吗?
INP.	当然 可以, 乔伊 不让 (你) 买 (它)?
REF.	Sure. Joey wouldn't let you buy it?
EXTE.	Of course. Sure, Joey won't get *it*?
JOIN.	Sure. Joey won't let *us* buy *one*?
+DIS.	Sure. Joey wouldn't let **you** buy **it**?
Non-Fixed Error	
PRE.	我 和 露西 只是 要 搬 到 对门。
INP.	我们 一 分手 (我) 就 搬 回去。
REF.	Once we broke up, I'll move back.
EXTE.	Once we broke up, *she*'ll move back.
JOIN.	Once we broke up, *we* moved back.
+DIS.	Once we broke up, *we*'ll move back.

Table 7: Example translations where pronouns in brackets are dropped in original inputs ("INP.") but labeled by humans according to references ("REF.") and previous sentence ("PRE."). We italicize some *mistranslated* errors and highlight the **correct** ones in bold.

overcome the data-level gap, Wang et al. (2016) proposed an automatic approach of ZP annotation by utilizing an alignment matrix from a large parallel data. By using the translation-oriented ZP corpus, they exploited different approaches to alleviate ZP problems for translation models (Wang et al., 2016, 2018a,b). Note that Wang et al. (2018b) also explored to address the problem of error propagation by jointly predicting ZP words given ZP position information. However, this method still relies an external model that predicting ZP positions at decoding time. Instead, this work proposes a unified model without any additional ZP annotations in decoding, thus release reliance on external ZP prediction in practice.

Discourse-Aware NMT Recent years, context-aware architecture has been well studied for NMT (Wang et al., 2017; Jean et al., 2017a; Tu et al., 2018). Wang et al. (2017) proposed hierarchical recurrent neural networks to summarize inter-sentential context from previous sentences and then integrate it into a standard NMT model with difference strategies. Jean et al. (2017a) introduced an additional set of an encoder and attention to encode and select part of the previous source sentence for generating each target word. Besides, Tu et al. (2018) proposed to augment NMT models with a cache-like memory network, which stores the translation history in terms of bilingual hidden representations at decoding steps of previous sentences. They also evaluated the above three models on different domains of data, showing that the hierarchical encoder performs comparable with the multi-attention model. More recently, some researchers began to investigate the effects of context-aware NMT on cross-lingual pronoun prediction (Jean et al., 2017b; Bawden et al., 2018; Voita et al., 2018). They mainly exploited general anaphora in non-pro-drop languages such as English⇒Russian.

6 Conclusion

In this work, we proposed a unified model to learn jointly predict and translate ZPs by leveraging multi-task learning. We also employed hierarchical neural networks to exploit discourse-level information for better ZP prediction. Experimental results on both Chinese⇒English and Japanese⇒English data show that the two proposed approaches accumulatively improve both the translation performance and ZP prediction accuracy. Our models also outperform the existing ZP translation models in previous work, and achieve a new state-of-the-art on the widely-used subtitle corpus. Manual evaluation confirms that the performance improvement comes from the alleviation of translation errors, which are mainly caused by subjective, objective as well as discourse-aware ZPs.

There are two potential extensions to our work. First, we will evaluate our method on other implication phenomena (or called unaligned words (Takeno et al., 2017)) such as tenses and article words for NMT. Second, we will investigate the impact of different context-aware models on ZP translation, including multi-attention (Jean et al., 2017b) and context-aware Transformer(Voita et al., 2018).

References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *NAACL*.

Chen Chen and Vincent Ng. 2013. Chinese zero pronoun resolution: Some recent advances. In *EMNLP*.

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 *EMNLP*.

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

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

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

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017a. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017b. Neural machine translation for cross-lingual pronoun prediction. In *Workshop on Discourse in Machine Translation*.

Fang Kong and Guodong Zhou. 2010. A tree kernel-based unified framework for chinese zero anaphora resolution. In *EMNLP*.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *WMT-MetricsMATR*.

Charles N Li and SA Thomson. 1979. Third-person pronouns and zero-anaphora in chinese discourse in discourse and syntax. *Syntax and Semantics*, 12.

Yang Liu and Qun Liu. 2010. Joint parsing and translation. In *COLING*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*.

Arum Park, Seunghee Lim, and Munpyo Hong. 2015. Zero object resolution in korean. In *PACLIC*.

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

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *CIKM*.

Hirotoshi Taira, Katsuhito Sudoh, and Masaaki Nagata. 2012. Zero pronoun resolution can improve the quality of J-E translation. In *Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Shunsuke Takeno, Masaaki Nagata, and Kazuhide Yamamoto. 2017. Controlling target features in neural machine translation via prefix constraints. In *The 4th Workshop on Asian Translation*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*.

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. *TACL*.

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

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. 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, Andy Way, and Qun Liu. 2017. 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 for dropped pronoun translation. In *NAACL*.

Bing Xiang, Xiaoqiang Luo, and Bowen Zhou. 2013. Enlisting the ghost: Modeling empty categories for machine translation. In *ACL*.

Xinyan Xiao, Yang Liu, Young-Sook Hwang, Qun Liu, and Shouxun Lin. 2010. Joint tokenization and translation. In *COLING*.

Qingyu Yin, Yu Zhang, Weinan Zhang, and Ting Liu. 2017. Chinese zero pronoun resolution with deep memory network. In *EMNLP*.

Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of chinese zero pronouns: A machine learning approach. In *EMNLP*.