Translating Pro-Drop Languages with Reconstruction Models

Longyue Wang
ADAPT Centre, Dublin City Univ.
longyue.wang@adaptcentre.ie

Zhaopeng Tu[*]
Tencent AI Lab
zptu@tencent.com

Shuming Shi
Tencent AI Lab
shumingshi@tencent.com

Tong Zhang
Tencent AI Lab
bradymzhang@tencent.com

Yvette Graham
ADAPT Centre, Dublin City Univ.
yvette.graham@adaptcentre.ie

Qun Liu
ADAPT Centre, Dublin City Univ.
qun.liu@adaptcentre.ie

Abstract

Pronouns are frequently omitted in pro-drop languages, such as Chinese, generally leading to significant challenges with respect to the production of complete translations. To date, very little attention has been paid to the dropped pronoun (DP) problem within neural machine translation (NMT). In this work, we propose a novel reconstruction-based approach to alleviating DP translation problems for NMT models. Firstly, DPs within all source sentences are automatically annotated with parallel information extracted from the bilingual training corpus. Next, the annotated source sentence is reconstructed from hidden representations in the NMT model. With auxiliary training objectives, in terms of reconstruction scores, the parameters associated with the NMT model are guided to produce enhanced hidden representations that are encouraged as much as possible to embed annotated DP information. Experimental results on both Chinese–English and Japanese–English dialogue translation tasks show that the proposed approach significantly and consistently improves translation performance over a strong NMT baseline, which is directly built on the training data annotated with DPs.

Introduction

In pro-drop languages, such as Chinese and Japanese, pronouns can be omitted from sentences when it is possible to infer the referent from the context. When translating sentences from a pro-drop language to a non-pro-drop language (*e.g.,* Chinese to English), machine translation systems generally fail to translate invisible dropped pronouns (DPs). This problem is especially severe in informal genres such as dialogues and conversation, where pronouns are more frequently omitted to make utterances more compact (Yang, Liu, and Xue 2015). For example, our analysis of a large Chinese–English dialogue corpus showed that around 26% of pronouns were dropped from the Chinese side of the corpus. This high proportion within informal genres shows the importance of addressing the challenge of translation of dropped pronouns.

Researchers have investigated methods of alleviating the DP problem for conventional Statistical Machine Transla-

Input	(它) 根本 没 那么 严重
Ref	**It** is not that bad
SMT	Wasn 't that bad
NMT	**It** 's not that bad
Input	这块 面包 很 美味 ! 你 烤 的 (它) 吗 ?
Ref	The bread is very tasty ! Did you bake **it** ?
SMT	This bread , delicious ! Did you bake ?
NMT	The bread is delicious ! Are you baked ?

Table 1: Examples of translating DPs where words in brackets are dropped pronouns that are invisible in decoding. NMT model's successes on translating simple dummy pronoun (upper panel), while fails on a more complicated one (bottom panel); SMT model fails in both cases.

tion (SMT) models showing promising results (Le Nagard and Koehn 2010; Xiang, Luo, and Zhou 2013; Wang et al. 2016a). Modeling DP translation for the more advanced Neural Machine Translation (NMT) models, however, has received substantially less attention, resulting in low performance in this respect even for state-of-the-art approaches. NMT models, due to their ability to capture semantic information with distributed representations, currently only manage to successfully translate some simple DPs, but still fail when translating anything more complex. Table 1 includes typical examples of when our strong baseline NMT system fails to accurately translate dropped pronouns. In this paper, we narrow the gap between correct DP translation for NMT models to improve translation quality for pro-drop languages with advanced models.

More specifically, we propose a novel reconstruction-based approach to alleviate DP problems for NMT. Firstly, we explicitly and automatically label DPs for each source sentence in the training corpus using alignment information from the parallel corpus (Wang et al. 2016a). Accordingly, each training instance is represented as a triple ($\mathbf{x}$, $\mathbf{y}$, $\hat{\mathbf{x}}$), where $\mathbf{x}$ and $\mathbf{y}$ are source and target sentences, and $\hat{\mathbf{x}}$ is the labelled source sentence. Next, we apply a standard encoder-decoder NMT model to translate $\mathbf{x}$, and obtain two sequences of hidden states from both encoder and

---

[*]Zhaopeng Tu is the corresponding author.

decoder. This is followed by introduction of an additional *reconstructor* (Tu et al. 2017b) to reconstruct back to the labelled source sentence $\hat{\mathbf{x}}$ with hidden states from either encoder or decoder, or both components. The central idea behind is to guide the corresponding hidden states to embed the recalled source-side DP information and subsequently to help the NMT model generate the missing pronouns with these enhanced hidden representations. To this end, the reconstructor produces a *reconstruction loss*, which measures how well the DP can be recalled and serves as an auxiliary training objective. Additionally, the likelihood score produced by the standard encoder-decoder measures the quality of general translation and the reconstruction score measures the quality of DP translation, and linear interpolation of these scores is employed as an overall score for a given translation.

Experiments on a large-scale Chinese–English corpus show that the proposed approach significantly improves translation performance by addressing the DP translation problem. Furthermore, when reconstruction is applied only in training, it improves parameter training by producing better hidden representations that embed the DP information. Results show improvement over a strong NMT baseline system of +1.35 BLEU points without any increase in decoding speed. When additionally applying reconstruction during testing, we obtain a further +1.06 BLEU point improvement with only a slight decrease in decoding speed of approximately 18%. Experiments for Japanese–English translation task show a significant improvement of 1.29 BLEU points, demonstrating the potential universality of the proposed approach across language pairs.

Contributions Our main contributions can be summarized as follows:

1. We show that although NMT models advance SMT models on translating pro-drop languages, there is still large room for improvement;

2. We introduce a reconstruction-based approach to improve dropped pronoun translation;

3. We release a large-scale bilingual dialogue corpus, which consists of 2.2M Chinese–English sentence pairs.[1]

Background
Pro-Drop Language Translation

A pro-drop language is a language in which certain classes of pronouns are omitted to make the sentence compact yet comprehensible when the identity of the pronouns can be inferred from the context. Since pronouns contain rich anaphora knowledge in discourse and the sentences in dialogue are generally short, DPs not only result in missing translations of pronouns, but also harm the sentence structure and even the semantics of output. Take the second case in Table 1 as an example, when the object pronoun "它" is dropped, the sentence is translated into "Are you baked?",

---

[1] Our released corpus is available at https://github.com/longyuewangdcu/tvsub.

Genres	Sents	ZH-Pro	EN-Pro	DP
Dialogue	2.15M	1.66M	2.26M	26.55%
Newswire	3.29M	2.27M	2.45M	7.35%

Table 2: Extent of DP in different genres. The *Dialogue* corpus consists of subtitles extracted from movie subtitle websites; The *Newswire* corpus is CWMT2013 news data.

while the correct translation should be "Did you bake it?". Such omissions may not be problematic for humans since they can easily recall missing pronouns from the context. They do, however, cause challenges for machine translation from a source pro-drop language to a target non-pro-drop language, since translation of such dropped pronouns generally fails.

As shown in Table 2, we analyzed two large Chinese–English corpora and found that around 26.55% of English pronouns can be dropped in the dialogue domain, while only 7.35% of pronouns were dropped in the newswire domain. DPs in formal text genres (*e.g.,* newswire) are not as common as those in informal genres (*e.g.,* dialogue), and the most frequently dropped pronouns in Chinese newswire is the third person singular 它 ("it") (Baran, Yang, and Xue 2012), which may not be crucial to translation performance. As the dropped pronoun phenomenon is more prevalent in informal genres, we test our method with respect to the dialogue domain.

Encoder-Decoder Based NMT

Neural machine translation (Sutskever, Vinyals, and Le 2014; Bahdanau, Cho, and Bengio 2015) has greatly advanced state-of-the-art within machine translation. Encoder-decoder architecture is now widely employed, where the encoder summarizes the source sentence $\mathbf{x} = x_1, x_2, \ldots, x_J$ into a sequence of hidden states $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_J\}$. Based on the encoder-side hidden states, the decoder generates the target sentence $\mathbf{y} = y_1, y_2, \ldots, y_I$ word by word with another sequence of decoder-side hidden states $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_I\}$:

$$P(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{I} P(y_i|y_{<i}, \mathbf{x}) = \prod_{i=1}^{I} g(y_{i-1}, \mathbf{s}_i, \mathbf{c}_i) \quad (1)$$

where $g(\cdot)$ is a softmax layer. The decoder hidden state $\mathbf{s}_i$ at step i is computed as

$$\mathbf{s}_i = f(y_{i-1}, \mathbf{s}_{i-1}, \mathbf{c}_i) \quad (2)$$

where $f(\cdot)$ is an activation function. $\mathbf{c}_i$ is a weighted sum of encoder hidden states $\mathbf{c}_t = \sum_{j=1}^{J} \alpha_{t,j} \mathbf{h}_j$, where $\alpha_{t,j}$ is the alignment probability calculated by an attention model (Bahdanau, Cho, and Bengio 2015; Luong, Pham, and Manning 2015). The parameters of the NMT model are trained to maximize the likelihood of a set of training examples $\{[\mathbf{x}^n, \mathbf{y}^n]\}_{n=1}^{N}$:

$$\mathcal{L}(\theta) = \arg\max_{\theta} \sum_{n=1}^{N} \log P(\mathbf{y}^n|\mathbf{x}^n; \theta) \quad (3)$$

System	Baseline	Oracle	△
SMT	30.16	35.26	+5.10
NMT	31.80	36.73	+4.93

Table 3: Translation performance improvement ("△") with manually labelled DPs ("Oracle").

Ideally, the hidden states (either encoder-side or decoder-side) should embed the missing DP information by learning the alignments between bilingual pronouns from the training corpus. In practice, however, complex DPs are still not translated correctly, as shown in Table 1. Table 3 shows empirical results to validate this assumption. We make the following two observations: (1) the NMT model indeed outperform SMT model when translating pro-drop languages; and (2) the performance of the NMT model can be further improved by improving translation of DPs. In this work, we propose to improve DP translation by guiding hidden states to embed the missing DP information.

Approach

In the following, we discuss methods of extending NMT models with a *reconstructor* to improve DP translation, which is inspired by "reconstruction" – a standard concept in auto-encoder (Bourlard and Kamp 1988; Vincent et al. 2010; Socher et al. 2011), and successfully applied to NMT models (Tu et al. 2017b) recently.

Architecture

Reconstructor The basic idea of our approach is to reconstruct the labelled source sentence from the latent representations of the NMT model and use the reconstruction score to measure how well the DPs can be recalled from latent representations. With the reconstruction score as an auxiliary training objective, we aim to encourage the latent representations to embed DP information, and thus recall the DP translation with enhanced representations.

The reconstructor reads a sequence of hidden states and the labelled source sentence, and outputs a reconstruction score. It employs an attention model (Bahdanau, Cho, and Bengio 2015; Luong, Pham, and Manning 2015) to reconstruct the labelled source sentence $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_{J'}\}$ word by word, which is conditioned on the input latent representations $\mathbf{v} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_T\}$. The reconstruction score is computed by

$$R(\hat{\mathbf{x}}|\mathbf{v}) = \prod_{j=1}^{J'} R(\hat{x}_j|\hat{x}_{<j}, \mathbf{v}) = \prod_{j=1}^{J'} g_r(\hat{x}_{j-1}, \hat{\mathbf{s}}_j, \hat{\mathbf{c}}_j) \quad (4)$$

where $\hat{\mathbf{s}}_j$ is the hidden state in the reconstructor, and computed by

$$\hat{\mathbf{s}}_j = f_r(\hat{x}_{j-1}, \hat{\mathbf{s}}_{j-1}, \hat{\mathbf{c}}_j) \quad (5)$$

Here $g_r(\cdot)$ and $f_r(\cdot)$ are respective softmax and activation functions for the reconstructor. The context vector $\hat{\mathbf{c}}_j$ is computed as a weighted sum of hidden states $\mathbf{v}$

$$\hat{\mathbf{c}}_j = \sum_{t=1}^{T} \hat{\alpha}_{j,t} \cdot \mathbf{v}_t \quad (6)$$
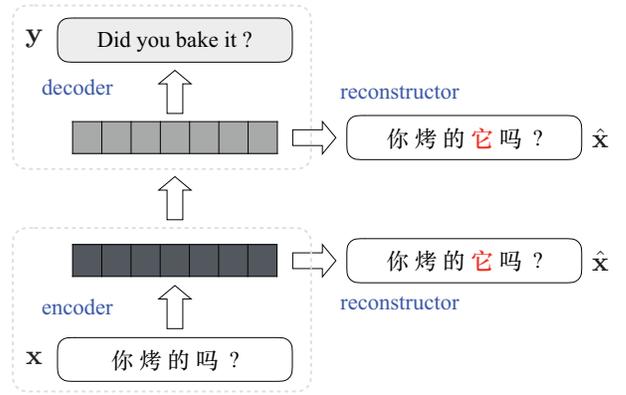

Figure 1: Architecture of reconstructor-augmented NMT. The two independent reconstructors reconstruct the labelled source sentence from hidden states in the encoder and decoder, respectively.

where the weight $\hat{\alpha}_{j,t}$ is calculated by an additional attention model. The parameters related to the attention model, $g_r(\cdot)$, and $f_r(\cdot)$ are independent of the standard NMT model. The labeled source words $\hat{\mathbf{x}}$ share the same word embeddings with the NMT encoder.

Reconstructor-Augmented NMT We augment the standard encoder-decoder based NMT model with the introduced reconstructor, as shown in Figure 1. The standard encoder-decoder reads the source sentence $\mathbf{x}$ and outputs its translation $\mathbf{y}$ along with the likelihood score. We introduce two independent reconstructors with their own parameters, each of which reconstructs the labelled source sentence $\hat{\mathbf{x}}$ from the encoder and decoder hidden states, respectively.

Note that we can append only one reconstructor to either the encoder or decoder:

- **(encoder-reconstructor)-decoder**: When adding a reconstructor to the encoder side only, we replace the standard encoder with an enhanced *auto-encoder*. In the case of auto-encoding, the encoder hidden states are not only used to summarize the original source sentence but also to embed the recalled DP information from the labelled source sentence.

- **encoder-(decoder-reconstructor)**: This is analogous to the framework proposed by Tu et al. (2017b), except that we reconstruct back to the labelled source sentence rather than the original one. It encourages the decoder hidden states to embed complete information from the source side, including the recalled DPs.

As seen, reconstructors applied on different sides of the corpus may capture different patterns of DP information, and using them together can encourage both the encoder and decoder to learn recalled DP information. Our approach is very much inspired by recent successes within question-answering, where a single information source is fed to multiple memory layers so that new evidence is captured in each layer and combined into subsequent layers (Sukhbaatar et al. 2015; Miller et al. 2016).

Data	$|S|$	$|W|$		$|P|$		$|V|$		$|L|$	
		Zh	En	Zh	En	Zh	En	Zh	En
Train	2.15M	12.1M	16.6M	1.66M	2.26M	151K	90.8K	5.63	7.71
Tune	1.09K	6.67K	9.25K	0.76K	1.03K	1.74K	1.35K	6.14	8.52
Test	1.15K	6.71K	9.49K	0.77K	0.96K	1.79K	1.39K	5.82	8.23

Table 4: Number of sentences ($|S|$), words ($|W|$), pronouns ($|P|$), vocabulary ($|V|$), and averaged sentence length ($|L|$) comprising the training, tuning and test corpora. K stands for thousands and M for millions.

Training and Testing

Training We train both the encoder-decoder and the introduced reconstructors together in a single end-to-end process. The two-reconstructor model (Figure 1) are described below (the other two individual models correspond to each part). The training objective can be revised as

$$J(\theta, \gamma, \psi) = \underset{\theta, \gamma, \psi}{\arg\max} \sum_{n=1}^{N} \left\{ \underbrace{\log P(\mathbf{y}^n | \mathbf{x}^n; \theta)}_{likelihood} \right.$$
$$\left. + \underbrace{\log R_{enc}(\hat{\mathbf{x}}^n | \mathbf{h}^n; \theta, \gamma)}_{enc\text{-}rec} + \underbrace{\log R_{dec}(\hat{\mathbf{x}}^n | \mathbf{s}^n; \theta, \psi)}_{dec\text{-}rec} \right\} \quad (7)$$

where θ is the parameter matrix in encoder-decoder, and γ and ψ are model parameters related to the *encoder-side reconstructor* ("enc-dec") and *decoder-side reconstructor* ("dec-rec") respectivley; $\mathbf{h}$ and $\mathbf{s}$ are encoder and decoder hidden states. The auxiliary reconstruction objectives (*e.g.,* $R_{enc}(\cdot)$ and $R_{dec}(\cdot)$) guide the related part of the parameter matrix θ to learn better latent representations, which are used to reconstruct the labelled source sentence.

Testing In testing, reconstruction can serve as a reranking technique to select a better translation from the k-best candidates generated by the decoder. Each translation candidate is assigned a likelihood score from the standard encoder-decoder, as well as reconstruction score(s) from the newly added reconstructor(s). Since the target sentence is invisible in testing, we employ a monolingual labelling model built on the training corpus to label DPs in the input sentence (Wang et al. 2016a).

When using reconstruction in testing, it requires external resources (*i.e.,* monolingual DP label tool) and more computations (*i.e.,* calculation of reconstruction scores). To reduce the dependency and cost, we can also employ a standard encoder-decoder model with better trained parameters so that the parameters can produce enhanced latent representations that embed DP information. Such information is invisible in the original input sentence but can be learned from the training data with similar context.

Experiments

Data

Experiments evaluate the method for translation of Chinese–English subtitles. More than two million sentence pairs were extracted from the subtitles of television episodes.[2] We pre-processed the extracted data using our in-house scripts (Wang et al. 2016b), including sentence boundary detection and bilingual sentence alignment etc. Finally, we obtained a high-quality corpus which keeps the discourse information. Table 4 lists the statistics of the corpus. Within the subtitle corpus, sentences are generally short and the Chinese side, as expected, contains many examples of dropped pronouns. We randomly select two complete television episodes as the tuning set, and another two episodes as the test set. We used case-insensitive 4-gram NIST BLEU metrics (Papineni et al. 2002) for evaluation, and *sign-test* (Collins, Koehn, and Kucerova 2005) to test for statistical significance.

DP Annotation

We follow Wang et al. (2016a) to automatically label DPs for training and test data. In the *training phase*, where the target sentence is available, we label DPs for the source sentence using alignment information. These labeled source sentences can be used to build a monolingual DP generator using NN, which is used to label test sentences since the target sentence is not available during the *testing phase*. The F1 scores of the two approaches on our data are 92.99% and 65.21%, respectively. After automatic labelling, the number of pronouns on the Chinese side in training, tuning and test data are 2.09M, 0.98K, 0.96K respectively, which is roughly consistent with pronoun frequency on the English side.

The usage of the labeled source sentences is two-fold:

1. *Baseline (+DPs)*: a stronger baseline system trained on the new parallel corpus (labelled source sentence, target sentence), which is evaluated on the new test sentences labelled by the monolingual DP generator.

2. *Our models*: the proposed models reconstruct hidden states back to the labelled source sentences.

For the source sentences that have no DPs, we use the original ones as labelled source sentences, otherwise we use the DP-labeled sentences.

Model

The baseline is our re-implemented attention-based NMT system, 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 baseline

---

[2]The data were crawled from the subtitle website http://www.zimuzu.tv.

Model	#Params	Speed		BLEU	
		Training	Decoding	Test	△
Baseline	86.7M	1.60K	2.61	31.80	– / –
Baseline (+DPs)	86.7M	1.59K	2.63	32.67†	+0.87 / –
+ enc-rec	+39.7M	0.71K	2.63	33.67†‡	+1.87 / +1.00
+ dec-rec	+34.1M	0.84K	2.18	33.48†‡	+1.68 / +0.81
+ enc-rec + dec-rec	+73.8M	0.57K	2.16	**35.08†‡**	**+3.28 / +2.41**
Multi-Source (Zoph and Knight 2016)	+20.7M	1.17K	1.27	32.81†	+1.01 / +0.14
Multi-Layer (Wu et al. 2016)	+75.1M	0.61K	2.42	33.36†	+1.56 / +0.69
Baseline (+DPs) + Enlarged Hidden Layer	+86.6M	0.68K	2.51	32.00†	+0.20 / -0.67

Table 5: Evaluation of translation performance for Chinese–English. "Baseline" is trained and evaluated on the original data, while "Baseline (+DPs)" is trained on the data labelled with DPs. "enc-rec" indicates encoder-side reconstructor and "dec-rec" denotes decoder-side reconstructor. Training speed is measured in words/second and decoding speed is measured in sentences/second with beam size being 10. The two numbers in the "△" column denote performance improvements over "Baseline" and "Baseline (+DPs)", respectively. "†" and "‡" indicate statistically significant difference ($p < 0.01$) from "Baseline" and "Baseline (+DPs)", respectively. All listed models except "Baseline" exploit the labelled source sentences.

models, we limited the source and target vocabularies to the most frequent 30K words in Chinese and English, covering approximately 97.2% and 99.3% of the words in the two languages, respectively. 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 word-embedding dimension 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 the tuning set.

The proposed model was implemented on top of the baseline model with the same settings where applicable. The hidden layer size in the reconstructor was 1,000. Following Tu et al. (2017b), we initialized the parameters of our models (*i.e.,* encoder and decoder, except those related to reconstructors) with the baseline model. We further trained all the parameters of our model for another 15 epochs.

Results and Discussion

Table 5 shows translation performances for Chinese–English. Clearly the proposed models significantly improve the translation quality in all cases, although there are still considerable differences among different variants.

Baselines The two baseline NMT models, one being trained and evaluated on the original bilingual data without any explicitly labelled DPs (*i.e.,* "Baseline"), while the other was trained and evaluated on the labelled data (*i.e.,* "Baseline (+DPs)"). As can be seen from the BLEU scores, the latter significantly outperforms the former, indicating that explicitly recalling translation of DPs helps produce better translations. Benefiting from the explicitly labelled DPs, the stronger baseline system is able to improve performance over the standard baseline system built on the original data where the pronouns are missing.

Parameters In terms of additional parameters introduced by the reconstruction models, both reconstructors introduce a large number of parameters. Beginning with the baseline model's 86.7M parameters, the encoder-side reconstructor adds 39.7M new parameters, while the decoder-side reconstructor adds a further 34.1M new parameters. Furthermore, adding reconstructors to both sides leads to additional 73.8M parameters. More parameters may capture more information, at the cost of posing difficulties to training.

Speed Although gains are made in terms of translation quality by introducing reconstruction, we need to consider the potential trade-off with respect to a possible increase in training and decoding times, due to the large number of newly introduced parameters resulting from the incorporation of reconstructors into the NMT model. When running on a single GPU device Tesla K80, the training speed of the baseline model is 1.60K target words per second, and this reduces to 0.57K words per second when reconstructors are added to both sides. In terms of decoding time trade-off, our most complex model only decreases decoding speed by 18%. We attribute this to the fact that no beam search is required for calculating reconstruction scores, which avoids the very costly data swap between GPU and CPU memories.

Translation Quality Clearly the proposed approach significantly improves the translation quality in all cases, although there are still considerable differences among the proposed variants. Introducing encoder-side and decoder-side reconstructors individually improves translation performance over "Baseline (+DPs)" by +1.0 and +0.8 BLEU points respectively. Combining them together achieves the best performance overall, which is +2.4 BLEU points better than the strong baseline model. This confirms our assumption that reconstructors applied to the source and target sides indeed capture different patterns for translating DPs.

Comparison to Other Work For the purpose of comparison, we reimplemented the multi-source model of Zoph and Knight (2016), which introduces an alternate encoder (shared parameters) and attention model (independent parameters) that take labelled sentences as an additional input source. This multi-source model significantly outperforms our "Baseline" model without labelled DP information, but only marginally outperform the "Baseline (+DPs)" that uses labelled DPs. One possible reason is that the two sources (*i.e.,* original input and labelled input sentences) are too similar to one another, making it difficult to distinguish them from labelled DPs.

Some may argue that the BLEU improvements are mainly due to the model parameter increase (*e.g.,* +73.8M) or deeper layers (*e.g.,* two reconstruction layers). To answer this concern, we compared the following two models:

- Multi-Layer (Wu et al. 2016): a system with three-layer encoder and three-layer decoder. The additional layers introduce 75.1M parameters, which is in the same scale with the proposed model (*i.e.,* 73.8M).

- Baseline (+DPs) + Enlarged Hidden Layer: a system with the same setting as "Baseline (+DPs)" except that layer size is 2100 instead of 1000. This variant introduces 86.6M parameters, which is even more than the most complicated variant of proposed models.

We found that the multi-layer model significantly outperforms its single-layer counterpart "Baseline (+DPs)", while significantly underperforms our best model (*i.e.,* 33.46 vs. 35.08). The "Baseline (+DPs)" system with enlarged hidden layer, however, does not achieve any improvement. This indicates that explicitly modeling DP translation is the key factor to the performance improvement.

Model	Test	△
Baseline (+DPs)	20.55	–
+ enc-rec + dec-rec	21.84	+ 1.29

Table 6: Evaluation of translation performance for Japanese–English.

Japanese–English Translation Task To validate the robustness of our approach on other pro-drop languages, we conducted experiments on Opensubtitle2016[3] data for the Japanese–English translation. We used the same settings as used in Chinese–English experiments, except that the vocabulary size is 20,001. As shown in Table 6, our model also significantly improves translation performance on the Japanese–English task, demonstrating the efficiency and potential universality of the proposed approach.

Analysis

We conducted extensive analyses for Chinese–English translation to better understand our model in terms of contribution of reconstruction from training and testing, effect of re-

constructed input, effect of DP labelling accuracy, and building the ability to handling long sentences.

Model	Test	△
Baseline	31.80	– / –
Baseline (+DPs)	32.67	+0.87 / –
+ enc-rec	33.67	+1.87 / +1.00
+ dec-rec	33.15	+1.35 / +0.48
+ enc-rec + dec-rec	**34.02**	**+2.22 / +1.35**

Table 7: Translation results when *reconstruction is used in training only while not used in testing*.

Contribution Analysis As mentioned previously, the effect of reconstruction is two-fold: (1) it improves the training of baseline parameters, which leads to better hidden representations that embed labelled DP information learned from the training data; and (2) it serves as a reranking metric in testing to measure the quality of DP translation.[4] Table 7 lists translation results when the reconstruction model is used in training only. Results show all variants to outperform the baseline models and applying reconstructors to both sides achieves the best performance overall. This is encouraging, since no extra resources nor computation are introduced to online decoding, making the approach highly practical, for example for translation in industry applications.

Model	Test	△
Baseline	31.80	– / –
Baseline (+DPs)	32.67	+0.87 / –
+ enc-rec	33.21	+1.41 / +0.54
+ dec-rec	33.08	+1.28 / +0.41
+ enc-rec + dec-rec	**33.25**	**+1.45 / +0.58**

Table 8: Translation results when hidden states are *reconstructed into the original source sentence* instead of the source sentence labelled with DPs.

Effect of Reconstruction Some researchers may argue that the proposed method acts much like dual learning (He et al. 2016a) and reconstruction (Tu et al. 2017b) especially when sentences have no DPs, which can benefit to the overall translation, not just only with respect to DPs. To investigate to what degree the improvements are indeed made by explicitly modeling DP translation, we examine the performance of variants which reconstruct hidden states back to the original input sentence instead of the source sentence labelled with DPs, as shown in Table 8. Note that the variant "+ dec-rec" in this setting is exactly the model proposed by Tu et al. (2017b). As seen, although the variants significantly outperforms "Baseline" model without using any DP

information, the absolute improvements are still worse than our proposed model that explicitly exploits DP information (*i.e.,* 1.45 vs. 3.28). This validates our hypothesis that explicitly modeling DP translation contributes most to the improvement.

Model	Automatic	Manual	$\triangle$
Baseline (+DPs)	32.67	36.73	+4.06
+ enc-rec	33.67	37.58	+3.91
+ dec-rec	33.48	37.23	+3.75
+ enc-rec + dec-rec	35.08	38.38	+3.30

Table 9: Translation performance gap ("$\triangle$") between manually ("Manual") and automatically ("Automatic") labelling DPs for input sentences in testing.

Effect of DP Labelling Accuracy   For each sentence in testing, the DPs are labelled automatically by a DP generator model, the accuracy of which is 65.21% measured in F1 score. The labelling errors may propagate to the NMT models, and have the potential to negatively affect translation performance. We investigate this using manual labelling and automatic labelling, as shown in Table 9. The analysis firstly shows that there still exists a significant gap in performance, and this could be improved by improving the accuracy of DP generator. Secondly, our models show a relatively smaller distance in performance from the oracle performance ("Manual"), indicating that the proposed approach is more robust to labelling errors.
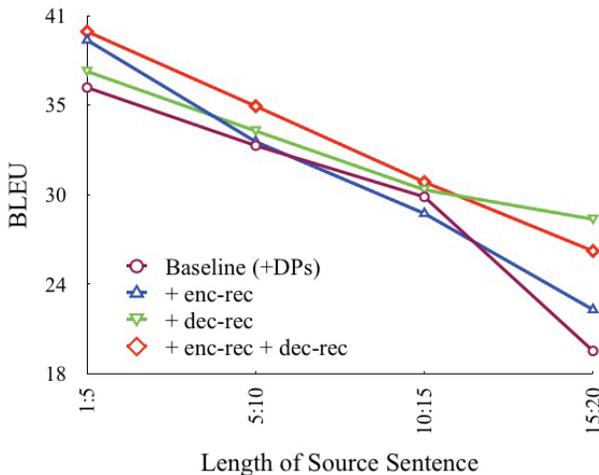

Figure 2: Performance of the generated translations with respect to the lengths of the source sentences.

Length Analysis   Following (Bahdanau, Cho, and Bengio 2015; Tu et al. 2016; 2017a), we group sentences of similar lengths together and compute the BLEU score for each group, as shown in Figure 2. The proposed models outperform the baseline for most span lengths, although there are still some notable differences. The improvement achieved by the source-side reconstructor is mainly for translation

of short sentences (*e.g.,* < 5), while that of the target-side reconstructor is mainly for translation of long sentences (*e.g.,* > 15). The reason is that (1) reconstruction can make encoder-side hidden states contain complete source information including DP information and subsequently good performance on short sentences, while at the same time, they cannot guarantee that all the information will be transferred to the decoder side (*i.e.,* relatively bad performance on long sentences); (2) similar to findings of (Tu et al. 2017b), the decoder-side reconstructor can make translation more adequate, which significantly alleviates inadequate translation problems for longer sentences. Combining them together can take advantage of both models, and thus the improvements are more substantial for all span lengths.

Model	Error	Sub.	Obj.	Dum.	All
BASE	Total	112	41	45	198
+ ENC	Fixed	51	22	28	101
	New	25	8	4	37
+ DEC	Fixed	57	21	17	95
	New	19	10	6	36
+ ENC + DEC	Fixed	50	34	33	117
	New	11	14	7	32

Table 10: Translation error statistics on different types of pronouns: subject ("Sub."), object ("Obj.") and dummy ("Dum.") pronouns. "BASE" denotes "Baseline (+DPs)", "+ ENC" denotes "+ enc-rec", "+ DEC" denotes "+ dec-rec" and "+ ENC + DEC" denotes "+ enc-rec + dec-rec".

Error Analysis   We investigate to what extent DP-related errors are fixed by the proposed models. We randomly select 500 sentences from the test set and count errors produced by the strong baseline model ("Total"), what proportion of these are fixed ("Fixed") or newly introduced ("New") by our approach, as shown in Table 10. All the proposed models can fix different kinds of DP problems, and the "+ ENC + DEC" variant achieves the best performance, which is consistent with the translation results reported above. The "+ ENC + DEC" model fixed 59.1% of the DP-related errors, while only introducing 16.2% of new errors. This confirms that our improvement in terms of automatic metric scores indeed comes from alleviating DP translation errors.

Among all types of pronouns, translation errors on object and dummy pronouns,[5] which can be usually inferred with intra-sentence context, are easy to be alleviated. In contrast, errors related to the subject of a given sentence are more difficult, since labelling dropped pronouns in such cases generally requires cross-sentence context. Table 11 shows three typical examples of successfully fixed, failed to fix, and newly introduced subjective-case pronouns.

---

[5]A dummy pronoun (*i.e.,* "it") is a pronoun used for syntax without explicit meaning. It is used in Germanic languages such as English but not in Pro-drop languages such as Chinese.

Fixed Error	
Input	等 我 搬进 来 (我) 可以 买 一台 泡泡机 吗 ？
Ref.	When I move in, can I get a bubble machine?
NMT	When I move in *to* buy a bubble machine.
Our	When I move in, can **I** buy a bubble machine?
Non-Fixed Error	
Input	(他) 是 个 训练营 ？
Ref.	It is a camp?
NMT	*He* was a camp?
Our	*He*'s a camp?
Newly Introduced Error	
Input	(我) 要 把 这 戒指 还给 你
Ref.	I need to give this ring back to you.
NMT	**I**'m gonna give you the ring back.
Our	*To* give it back to you.

Table 11: Example translations where subjective-case pronouns in brackets are dropped in original input but labeled by DP generator. We italicize some *mis-translated* errors and highlight the **correct** ones in bold.

Related Work

DP Translation for SMT Previous research has investigated DP translation for SMT. For example, Chung and Gildea (2010) examined the effects of empty category (including DPs) on MT with various methods. This work showed improvements in terms of translation performance despite the automatic prediction of empty category not being highly accurate. Taira, Sudoh, and Nagata (2012) analyzed the Japanese-to-English translation by inserting DPs into input sentences using simple rule-based methods, achieving marginal improvements. More recently, Wang et al. (2016a) proposed labelling DPs using parallel information of training data, and obtained promising results in SMT. Wang et al. (2017b) also extend the SMT-based DP translation method on Japanese–English translation task. Inspired by these previous successes, this paper is an early attempt to learn to tackle DP translation for NMT models.

Representation Learning with Reconstruction Reconstruction is a standard concept in auto-encoder, that guides towards learning representations that captures the underlying explanatory factors for the observed input (Bourlard and Kamp 1988; Vincent et al. 2010). An auto-encoder model consists of an encoding function to compute a representation from an input, and a decoding function to reconstruct the input from the representation. The parameters involved in the two functions are trained to maximize the *reconstruction score*, which measures the similarity between the original input and reconstructed input. Inspired by the concept of *reconstruction*, Tu et al. (2017b) proposed guiding decoder hidden states to embed complete source information by reconstructing the hidden states back to the original source sentence. Our approach differs at: (1) we introduced not only decoder-side reconstructor but also encoder-side reconstructor to learn enhanced hidden states of both encoder and decoder; and (2) we guide the hidden states to embed complete source information as well as the labelled DP information.

Multiple Sources for NMT Recently, it was shown that NMT can be improved by feeding auxiliary information sources beyond the original input sentence. The additional sources can be in various forms, such as parallel sentences in other languages (Dong et al. 2015; Zoph and Knight 2016), cross-sentence contexts (Wang et al. 2017a; Jean et al. 2017; Tu et al. 2018), generation recommendations from other translation models (He et al. 2016b; Wang et al. 2017c; Gu et al. 2017; Wang et al. 2017d), syntax information (Li et al. 2017; Zhou et al. 2017). Along the same direction, we provide complementary information in terms of source sentences labelled with DPs.

Conclusion and Future Work

This paper is an early attempt to model DP translation for NMT systems. Hidden states are guided in both the encoder and decoder to embed the DP information by reconstructing them back to the source sentence labelled with DPs. The effect of reconstruction model is two-fold: (1) it improves parameter training for producing better latent representations; and (2) it measures the quality of DP translation, which is combined with likelihood to better measure the overall quality of translations. Quantity and quality analyses show that the proposed approach significantly improves translation performance across language pairs, and can be further improved by developing better DP labelling models.

In future work we plan to validate the effectiveness of our approach on other text genres with different prevalence of DPs. For example, in formal text genres (*e.g.,* newswire), DPs are not as common as in the informal text genres, and the most frequently dropped pronouns in Chinese newswire is the third person singular "它" ("*it*") (Baran, Yang, and Xue 2012), which may not be crucial to translation performance.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*, 1–15.

Baran, E.; Yang, Y.; and Xue, N. 2012. Annotating dropped pronouns in chinese newswire text. In *LREC 2012*, 2795–2799.

Bourlard, H., and Kamp, Y. 1988. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics* 59(4-5):291–294.

Chung, T., and Gildea, D. 2010. Effects of empty categories on machine translation. In *EMNLP 2010*, 636–645.

Collins, M.; Koehn, P.; and Kucerova, I. 2005. Clause restructuring for statistical machine translation. In *ACL 2005*, 531–540.

Dong, D.; Wu, H.; He, W.; Yu, D.; and Wang, H. 2015. Multi-task learning for multiple language translation. In *ACL-IJCNLP 2015*, 1723–1732.

Gu, J.; Wang, Y.; Cho, K.; and Li, V. O. 2017. Search engine guided non-parametric neural machine translation. *arXiv preprint arXiv:1705.07267*.

He, D.; Xia, Y.; Qin, T.; Wang, L.; Yu, N.; Liu, T.; and Ma, W.-Y. 2016a. Dual learning for machine translation. In *NIPS 2016*, 820–828.

He, W.; He, Z.; Wu, H.; and Wang, H. 2016b. Improved neural machine translation with smt features. In *AAAI 2016*, 151–157.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Jean, S.; Lauly, S.; Firat, O.; and Cho, K. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Le Nagard, R., and Koehn, P. 2010. Aiding pronoun translation with co-reference resolution. In *WMT-MetricsMATR 2010*, 252–261.

Li, J.; Xiong, D.; Tu, Z.; Zhu, M.; Zhang, M.; and Zhou, G. 2017. Modeling source syntax for neural machine translation. In *ACL 2017*, 688–697.

Luong, T.; Pham, H.; and Manning, D. C. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP 2015*, 1412–1421.

Miller, A.; Fisch, A.; Dodge, J.; Karimi, A.-H.; Bordes, A.; and Weston, J. 2016. Key-value memory networks for directly reading documents. In *EMNLP 2016*, 1400–1409.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL 2002*, 311–318.

Socher, R.; Pennington, J.; Huang, E. H.; Ng, A. Y.; and Manning, C. D. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *EMNLP 2011*, 151–161.

Sukhbaatar, S.; Szlam, A.; Weston, J.; and Fergus, R. 2015. End-to-end memory networks. In *NIPS 2015*, 2440–2448.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS 2014*, 3104–3112.

Taira, H.; Sudoh, K.; and Nagata, M. 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*, 111–118.

Tu, Z.; Lu, Z.; Liu, Y.; Liu, X.; and Li, H. 2016. Modeling Coverage for Neural Machine Translation. In *ACL 2016*, 76–85.

Tu, Z.; Liu, Y.; Lu, Z.; Liu, X.; and Li, H. 2017a. Context gates for neural machine translation. In *TACL 2017*, 87–99.

Tu, Z.; Liu, Y.; Shang, L.; Liu, X.; and Li, H. 2017b. Neural machine translation with reconstruction. In *AAAI 2017*, 3097–3103.

Tu, Z.; Liu, Y.; Shi, S.; and Zhang, T. 2018. Learning to remember translation history with a continuous cache. In *TACL 2018*.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; and Manzagol, P.-A. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* 11(Dec):3371–3408.

Wang, L.; Tu, Z.; Zhang, X.; Li, H.; Way, A.; and Liu, Q. 2016a. A novel approach for dropped pronoun translation. In *NAACL 2016*, 983–993.

Wang, L.; Zhang, X.; Tu, Z.; Way, A.; and Liu, Q. 2016b. The automatic construction of discourse corpus for dialogue translation. In *LREC 2016*.

Wang, L.; Tu, Z.; Way, A.; and Liu, Q. 2017a. Exploiting cross-sentence context for neural machine translation. In *EMNLP 2017*, 2816–2821.

Wang, L.; Tu, Z.; Zhang, X.; Liu, S.; Li, H.; Way, A.; and Liu, Q. 2017b. A novel and robust approach for pro-drop language translation. *Machine Translation* 31(1):65–87.

Wang, X.; Lu, Z.; Tu, Z.; Li, H.; Xiong, D.; and Zhang, M. 2017c. Neural machine translation advised by statistical machine translation. In *AAAI 2017*, 3330–3336.

Wang, X.; Tu, Z.; Xiong, D.; and Zhang, M. 2017d. Translating phrases in neural machine translation. In *EMNLP 2017*, 1432–1442.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xiang, B.; Luo, X.; and Zhou, B. 2013. Enlisting the ghost: Modeling empty categories for machine translation. In *ACL 2013*, 822–831.

Yang, Y.; Liu, Y.; and Xue, N. 2015. Recovering dropped pronouns from chinese text messages. In *ACL 2015*, 309–313.

Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhou, H.; Tu, Z.; Huang, S.; Liu, X.; Li, H.; and Chen, J. 2017. Chunk-based bi-scale decoder for neural machine translation. In *ACL 2017*, 580–586.

Zoph, B., and Knight, K. 2016. Multi-source neural translation. In *NAACL 2016*, 30–34.