Learning to Jointly Translate and Predict Dropped Pronouns with a Shared Reconstruction Mechanism

Longyue Wang
Tencent AI Lab
vinnylywang@tencent.com

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

Andy Way
Dublin City University
andy.way@adaptcentre.ie

Qun Liu
Huawei Noah's Ark Lab
qun.liu@huawei.com

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. Recently, Wang et al. (2018) proposed a novel reconstruction-based approach to alleviating dropped pronoun (DP) translation problems for neural machine translation models. In this work, we improve the original model from two perspectives. First, we employ a shared reconstructor to better exploit encoder and decoder representations. Second, we jointly learn to translate and predict DPs in an end-to-end manner, to avoid the errors propagated from an external DP prediction model. Experimental results show that our approach significantly improves both translation performance and DP prediction accuracy.

1 Introduction

Pronouns are important in natural languages as they imply rich discourse information. However, in pro-drop languages such as Chinese and Japanese, pronouns are frequently omitted when their referents can be pragmatically inferred from the context. When translating sentences from a pro-drop language into a non-pro-drop language (*e.g.* Chinese-to-English), translation models generally fail to translate invisible dropped pronouns (DPs). This phenomenon leads to various translation problems in terms of completeness, syntax and even semantics of translations. A number of approaches have been investigated for DP translation (Le Nagard and Koehn, 2010; Xiang et al., 2013; Wang et al., 2016, 2018).

Wang et al. (2018) is a pioneering work to model DP translation for neural machine translation (NMT) models. They employ two *separate* reconstructors (Tu et al., 2017) to respectively reconstruct encoder and decoder representations back to the DP-annotated source sentence. The annotation of DP is provided by an *external* prediction model, which is trained on the parallel corpus using automatically learned alignment information (Wang et al., 2016). Although this model achieved significant improvements, there nonetheless exist two drawbacks: 1) there is no interaction between the two separate reconstructors, which misses the opportunity to exploit useful relations between encoder and decoder representations; and 2) the external DP prediction model only has an accuracy of 66% in F1-score, which propagates numerous errors to the translation model.

In this work, we propose to improve the original model from two perspectives. First, we use a *shared* reconstructor to read hidden states from both encoder and decoder. Second, we integrate a DP predictor into NMT to *jointly* learn to translate and predict DPs. Incorporating these as two auxiliary loss terms can guide both the encoder and decoder states to learn critical information relevant to DPs. Experimental results on a large-scale Chinese–English subtitle corpus show that the two modifications can accumulatively improve translation performance, and the best result is +1.5 BLEU points better than that reported by Wang et al. (2018). In addition, the jointly learned DP prediction model significantly outperforms its external counterpart by 9% in F1-score.

2 Background

As shown in Figure 1, Wang et al. (2018) introduced two independent reconstructors with their own parameters, which reconstruct the DP-annotated source sentence from the encoder and decoder hidden states, respectively. The central

[*] Zhaopeng Tu is the corresponding author of the paper. This work was conducted when Longyue Wang was studying and Qun Liu was working at the ADAPT Centre in the School of Computing at Dublin City University.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2997–3002
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

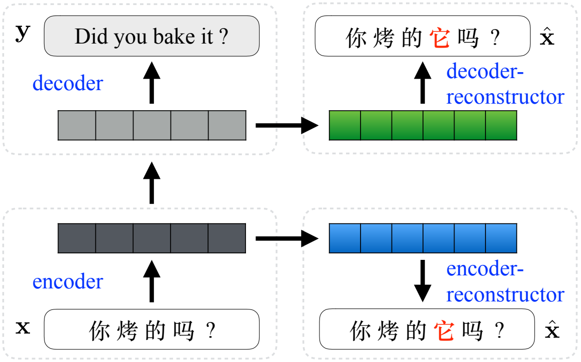


Figure 1: Architecture of separate reconstructors.

Prediction	F1-score	Example
DP Position	88%	你烤的#DP#吗？
DP Words	66%	你烤的它吗？

Table 1: Evaluation of external models on predicting the positions of DPs ("DP Position") and the exact words of DP ("DP Words").

idea underpinning their approach 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.

The DPs can be automatically annotated for training and test data using two different strategies (Wang et al., 2016). In the *training phase*, where the target sentence is available, we annotate DPs for the source sentence using alignment information. These annotated source sentences can be used to build a neural-based DP predictor, which can be used to annotate test sentences since the target sentence is not available during the *testing phase*. As shown in Table 1, Wang et al. (2016, 2018) explored to predict the exact DP words[1], the accuracy of which is only 66% in F1-score. By analyzing the translation outputs, we found that 16.2% of errors are newly introduced and caused by errors from the DP predictor. Fortunately, the accuracy of predicting DP positions (DPPs) is much higher, which provides the chance to alleviate the error propagation problem. Intuitively, we can learn to generate DPs at the predicted positions using a jointly trained DP predictor, which is fed with informative representations in the reconstructor.

3 Approach

3.1 Shared Reconstructor

Recent work shows that NMT models can benefit from sharing a component across different tasks and languages. Taking multi-language translation as an example, Firat et al. (2016) share an attention model across languages while Dong et al. (2015) share an encoder. Our work is most similar to the work of Zoph and Knight (2016) and Anastasopoulos and Chiang (2018), which share a decoder and two separate attention models to read from two different sources. In contrast, we share information at the level of reconstructed frames.

The architectures of our proposed shared reconstruction model are shown in Figure 2(a). Formally, the reconstructor reads from both the encoder and decoder hidden states, as well as the DP-annotated source sentence, and outputs a reconstruction score. It uses two separate attention models to reconstruct the annotated source sentence $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_T\}$ word by word, 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, and computed by Equation (1):

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

Here $g_r(\cdot)$ and $f_r(\cdot)$ are respectively 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 $\mathbf{h}^{dec}$, respectively, as in Equation (2) and (3):

$$\hat{\mathbf{c}}_t^{enc} = \textstyle\sum_{j=1}^{J} \hat{\alpha}_{t,j}^{enc} \cdot \mathbf{h}_j^{enc} \qquad (2)$$

$$\hat{\mathbf{c}}_t^{dec} = \textstyle\sum_{i=1}^{I} \hat{\alpha}_{t,i}^{dec} \cdot \mathbf{h}_i^{dec} \qquad (3)$$

Note that the weights $\hat{\alpha}^{enc}$ and $\hat{\alpha}^{dec}$ are calculated by two separate attention models. We propose two attention strategies which differ as to whether the two attention models have interactions or not.

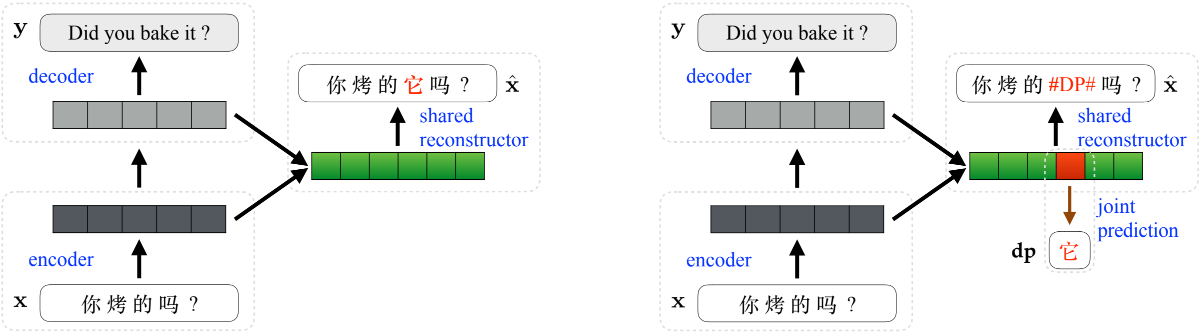
Independent Attention calculates the two weight matrices independently, as in Equation (4) and (5):

$$\hat{\alpha}^{enc} = \text{ATT}_{enc}(\hat{x}_{t-1}, \mathbf{h}_{t-1}^{rec}, \mathbf{h}^{enc}) \qquad (4)$$

$$\hat{\alpha}^{dec} = \text{ATT}_{dec}(\hat{x}_{t-1}, \mathbf{h}_{t-1}^{rec}, \mathbf{h}^{dec}) \qquad (5)$$

where $\text{ATT}_{enc}(\cdot)$ and $\text{ATT}_{dec}(\cdot)$ are two separate attention models with their own parameters.

[1] Unless otherwise indicated, in the paper, the terms "DP" and "DP word" are identical.



(a) Shared reconstructor.

(b) Shared reconstructor with joint prediction.

Figure 2: Model architectures in which the words in red are automatically annotated DPs and DPPs.

Interactive Attention feeds the context vector produced by one attention model to another attention model. The intuition behind this is that the interaction between two attention models can lead to a better exploitation of the encoder and decoder representations. As the interactive attention is directional, we have two options (Equation (6) and (7)) which modify either $\textsc{Att}_{enc}(\cdot)$ or $\textsc{Att}_{dec}(\cdot)$ while leaving the other one unchanged:

- *enc→dec*:

$$\hat{\alpha}^{dec} = \textsc{Att}_{dec}(\hat{x}_{t-1}, \mathbf{h}^{rec}_{t-1}, \mathbf{h}^{dec}, \hat{\mathbf{c}}^{enc}_{t}) \qquad (6)$$

- *dec→enc*:

$$\hat{\alpha}^{enc} = \textsc{Att}_{enc}(\hat{x}_{t-1}, \mathbf{h}^{rec}_{t-1}, \mathbf{h}^{enc}, \hat{\mathbf{c}}^{dec}_{t}) \qquad (7)$$

3.2 Joint Prediction of Dropped Pronouns

Inspired by recent successes of multi-task learning (Dong et al., 2015; Luong et al., 2016), we propose to jointly learn to translate and predict DPs (as shown in Figure 2(b)). To ease the learning difficulty, we leverage the information of DPPs predicted by an external model, which can achieve an accuracy of 88% in F1-score. Accordingly, we transform the original DP prediction problem to DP word generation given the pre-predicted DP positions. Since the DPP-annotated source sentence serves as the reconstructed input, we introduce an additional *DP-generation loss*, which measures how well the DP is generated from the corresponding hidden state in the reconstructor.

Let $\mathbf{dp} = \{dp_1, dp_2, \ldots, dp_D\}$ be the list of DPs in the annotated source sentence, and $\mathbf{h}^{rec} = \{\mathbf{h}^{rec}_1, \mathbf{h}^{rec}_2, \ldots, \mathbf{h}^{rec}_D\}$ be the corresponding hidden states in the reconstructor. The generation probability is computed by

$$\begin{aligned} P(\mathbf{dp}|\mathbf{h}^{rec}) &= \prod_{d=1}^{D} P(dp_d|\mathbf{h}^{rec}_d) \\ &= \prod_{d=1}^{D} g_p(dp_d|\mathbf{h}^{rec}_d) \end{aligned} \qquad (8)$$

where $g_p(\cdot)$ is softmax for the DP predictor.

3.3 Training and Testing

We train both the encoder-decoder and the shared reconstructors together in a single end-to-end process, and the training objective is

$$\begin{aligned} J(\theta,\gamma,\psi) = \arg\max_{\theta,\gamma,\psi} \Big\{ & \underbrace{\log L(\mathbf{y}|\mathbf{x};\theta)}_{\textit{likelihood}} \\ & + \underbrace{\log R(\hat{\mathbf{x}}|\mathbf{h}^{enc},\mathbf{h}^{dec};\theta,\gamma)}_{\textit{reconstruction}} \\ & + \underbrace{\log P(\mathbf{dp}|\hat{\mathbf{h}}^{rec};\theta,\gamma,\psi)}_{\textit{prediction}} \Big\} \end{aligned} \qquad (9)$$

where $\{\theta, \gamma, \psi\}$ are respectively the parameters associated with the encoder-decoder, shared reconstructor and the DP prediction model. The auxiliary reconstruction objective $R(\cdot)$ guides the related part of the parameter matrix θ to learn better latent representations, which are used to reconstruct the DPP-annotated source sentence. The auxiliary prediction loss $P(\cdot)$ guides the related part of both the encoder-decoder and the reconstructor to learn better latent representations, which are used to predict the DPs in the source sentence.

Following Tu et al. (2017) and Wang et al. (2018), we use the reconstruction score

#	Model	#Params	Speed		BLEU
			Train	Decode	
	Existing system (Wang et al., 2018)				
1	Baseline	86.7M	1.60K	15.23	31.80
2	Baseline (+DPs)	86.7M	1.59K	15.20	32.67
3	Separate-Recs⇒(+DPs)	+73.8M	0.57K	12.00	35.08
	Our system				
4	Baseline (+DPPs)	86.7M	1.54K	15.19	33.18
5	Shared-Rec$_{independent}$⇒(+DPPs)	+86.6M	0.52K	11.87	$35.27^{\dagger\ddagger}$
6	Shared-Rec$_{independent}$⇒(+DPPs) + joint prediction	+87.9M	0.51K	11.88	$35.88^{\dagger\ddagger}$
7	Shared-Rec$_{enc \rightarrow dec}$⇒(+DPPs) + joint prediction	+91.9M	0.48K	11.84	**36.53**†‡
8	Shared-Rec$_{dec \rightarrow enc}$⇒(+DPPs) + joint prediction	+89.9M	0.49K	11.85	$35.99^{\dagger\ddagger}$

Table 2: Evaluation of translation performance for Chinese–English. “Baseline” is trained and evaluated on the original data, while “Baseline (+DPs)” and “Baseline (+DPPs)” are trained on the data annotated with DPs and DPPs, respectively. Training and decoding (beam size is 10) speeds are measured in words/second. “†” and “‡” indicate statistically significant difference ($p < 0.01$) from “Baseline (+DDPs)” and “Separate-Recs⇒(+DPs)”, respectively.

as a reranking technique to select the best translation candidate from the generated n-best list at testing time. Different from Wang et al. (2018), we reconstruct DPP-annotated source sentence, which is predicted by an external model.

4 Experiment

4.1 Setup

To compare our work with the results reported by previous work (Wang et al., 2018), we conducted experiments on their released Chinese⇒English TV Subtitle corpus.[2] The training, validation, and test sets contain 2.15M, 1.09K, and 1.15K sentence pairs, respectively. We used case-insensitive 4-gram NIST BLEU metrics (Papineni et al., 2002) for evaluation, and *sign-test* (Collins et al., 2005) to test for statistical significance.

We implemented our models on the code repository released by Wang et al. (2018).[3] We used the same configurations (*e.g.* vocabulary size = 30K, hidden size = 1000) and reproduced their reported results. It should be emphasized that we did not use the pre-train strategy as done in Wang et al. (2018), since we found training from scratch achieved a better performance in the shared reconstructor setting.

4.2 Results

Table 2 shows the translation results. It is clear that the proposed models significantly outperform the baselines in all cases, although there are considerable differences among different variations.

Baselines (Rows 1-4): The three baselines (Rows 1, 2, and 4) differ regarding the training data used. “Separate-Recs⇒(+DPs)” (Row 3) is the best model reported in Wang et al. (2018), which we employed as another strong baseline. The baseline trained on the DPP-annotated data (“Baseline (+DPPs)”, Row 4) outperforms the other two counterparts, indicating that the error propagation problem does affect the performance of translating DPs. It suggests the necessity of jointly learning to translate and predict DPs.

Our Models (Rows 5-8): Using our shared reconstructor (Row 5) not only outperforms the corresponding baseline (Row 4), but also surpasses its separate reconstructor counterpart (Row 3). Introducing a joint prediction objective (Row 6) can achieve a further improvement of +0.61 BLEU points. These results verify that shared reconstructor and jointly predicting DPs can accumulatively improve translation performance.

Among the variations of shared reconstructors (Rows 6-8), we found that an interaction attention from encoder to decoder (Row 7) achieves the best performance, which is +3.45 BLEU points better than our baseline (Row 4) and +1.45 BLEU points

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

[3] https://github.com/tuzhaopeng/nmt

better than the best result reported by Wang et al. (2018) (Row 3). We attribute the superior performance of "Shared-Rec$_{enc\rightarrow dec}$" to the fact that the attention context over encoder representations embeds useful DP information, which can help to better attend to the representations of the corresponding pronouns in the decoder side. Similar to Wang et al. (2018), the proposed approach improves BLEU scores at the cost of decreased training and decoding speed, which is due to the large number of newly introduced parameters resulting from the incorporation of reconstructors into the NMT model.

4.3 Analysis

Models	Precision	Recall	F1-score
External	0.67	0.65	0.66
Joint	0.74	0.76	0.75

Table 3: Evaluation of DP prediction accuracy. "External" model is *separately* trained on DP-annotated data with external neural methods (Wang et al., 2016), while "Joint" model is *jointly* trained with the NMT model (Section 3.2).

DP Prediction Accuracy As shown in Table 3, the jointly learned model significantly outperforms the external one by 9% in F1-score. We attribute this to the useful contextual information embedded in the reconstructor representations, which are used to generate the exact DP words.

Model	Test	△
Baseline (+DPPs)	33.18	–
Separate-Recs (+DPs)	34.02	+0.84
Shared-Rec (+DPPs)	**34.80**	**+1.62**

Table 4: Translation results when *reconstruction is used in training only while not used in testing*.

Contribution Analysis Table 4 lists translation results when the reconstruction model is used in training only. We can see that the proposed model outperforms both the strong baseline and the best model reported in Wang et al. (2018). This is encouraging since no extra resources and computation are introduced to online decoding, which makes the approach highly practical, for example for translation in industry applications.

Model	Auto.	Man.	△
Seperate-Recs (+DPs)	35.08	38.38	+3.30
Shared-Rec (+DPPs)	36.53	38.94	+2.41

Table 5: Translation performance gap ("△") between manually ("Man.") and automatically ("Auto.") labelling DPs/DPPs for input sentences in testing.

Effect of DPP Labelling Accuracy For each sentence in testing, the DPs and DPPs are labelled automatically by two separate external prediction models, the accuracy of which are respectively 66% and 88% measured in F1 score. We investigate the best performance the models can achieve with manual labelling, which can be regarded as an "Oracle", as shown in Table 5. As seen, there still exists a significant gap in performance, and this could be improved by improving the accuracy of our DPP generator. In addition, our models show a relatively smaller distance in performance from the oracle performance ("Man"), indicating that the error propagation problem is alleviated to some extent.

5 Conclusion

In this paper, we proposed effective approaches of translating DPs with NMT models: *shared* reconstructor and *jointly* learning to translate and predict DPs. Through experiments we verified that 1) shared reconstruction is helpful to share knowledge between the encoder and decoder; and 2) joint learning of the DP prediction model indeed alleviates the error propagation problem by improving prediction accuracy. The two approaches accumulatively improve translation performance. The method is not restricted to the DP translation task and could potentially be applied to other sequence generation problems where additional source-side information could be incorporated.

In future work we plan to: 1) build a fully end-to-end NMT model for DP translation, which does not depend on any external component (*i.e.* DPP predictor); 2) exploit cross-sentence context (Wang et al., 2017) to further improve DP translation; 3) investigate a new research strand that adapts our model in an inverse translation direction by learning to drop pronouns instead of recovering DPs.

Acknowledgments

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. We thank the anonymous reviewers for their insightful comments.

References

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 82–91, New Orleans, Louisiana, USA.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 531–540, Ann Arbor, Michigan, USA.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1723–1732, Beijing, China.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California, USA.

Ronan Le Nagard and Philipp Koehn. 2010. Aiding pronoun translation with co-reference resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2965–2977, Santa Fe, New Mexico, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3097–3103, San Francisco, California, USA.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. Translating pro-drop languages with reconstruction models. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 4937–4945, New Orleans, Louisiana, USA.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2816–2821, Copenhagen, Denmark.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016. A novel approach for dropped pronoun translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 983–993, San Diego, California, USA.

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.

Barret Zoph and Knight Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California.