DROPPED PRONOUN GENERATION FOR DIALOGUE MACHINE TRANSLATION

Longyue Wang⋆ *Xiaojun Zhang*⋆ *Zhaopeng Tu*† *Hang Li*† *Qun Liu*⋆

⋆ADAPT Centre, School of Computing, Dublin City University
†Noah's Ark Lab, Huawei Technologies
⋆{lwang,xzhang,qliu}@computing.dcu.ie
†{tu.zhaopeng,hangli.hl}@huawei.com

ABSTRACT

Dropped pronoun (DP) is a common problem in dialogue machine translation, in which pronouns are frequently dropped in the source sentence and thus are missing in its translation. In response to this problem, we propose a novel approach to improve the translation of DPs for dialogue machine translation. Firstly, we build a training data for DP generation, in which the DPs are automatically added according to the alignment information from a parallel corpus. Then we model the DP generation problem as a sequence labelling task, and develop a generation model based on recurrent neural networks and language models. Finally, we apply the DP generator to machine translation task by completing the source sentences with the missing pronouns. Experimental results show that our approach achieves a significant improvement of 1.7 BLEU points by recalling possible DPs in the source sentences.

Index Terms— Machine Translation, Dialogue, Dropped Pronoun.

1. INTRODUCTION

A pro-drop language is a language in which certain classes of pronouns are omitted. For example, Chinese is a pro-drop language in which the nominal is frequently missing [1], while English is a not pro-drop language [2]. In the discourse translation from Chinese into English, pronouns as the anaphors are frequently dropped in the source language. Take Figure 1 as an example, the Chinese objects 它 (”*it*”), 我 (”*I*”), and 你 (”*you*”) are eliminated in the sentences. We further validate this finding by analysing a large Chinese-English dialogue corpus.[1] We find that there are 7.8M Chinese pronouns and 9.2M English pronouns, which denotes that 1.4M Chinese pronouns are missing in this parallel corpus.

These omissions may not be problems for humans as people can easily recall the missing pronouns from the context. However, this does not hold in statistical machine translation (SMT) scenario, since most missing pronouns fail to be translated appropriately, even with the state-of-the-art translation models. While this problem has motivated significant research in the field of natural language processing (NLP), the integration of those outputs (e.g., co-reference resolution (CR)) into machine translation has been lacking, even the recent wave of work on SMT has not touched this obstacle. In this paper, we propose a novel approach to crack the barrier of DP translation by recalling possible DPs in the source sentences. To this end,

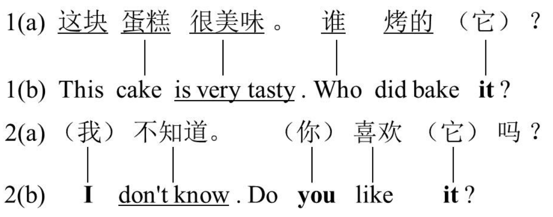
1(a) 这块 蛋糕 很美味 。 谁 烤的 （它）？
1(b) This cake is very tasty . Who did bake **it** ?
2(a) （我） 不知道。 （你） 喜欢 （它） 吗 ?
2(b) **I** don't know . Do **you** like **it** ?

Fig. 1. Examples of dropped pronouns in a parallel dialogue corpus. The Chinese pronouns in the brackets are missing.

1. Given a parallel corpus, we use the source side as the training data for DP generation, which is automatically annotated with DPs by projecting aligned pronouns at the target side. We believe this is the first work on recovering DPs by using parallel corpus.

2. With the DP-inserted corpus, we propose a two-phase DP generation approach. First, we employ a sequence labelling model based on recurrent neural networks (RNNs) to predict the positions and the categories of DPs in the source sentences. Second, we use another n-gram language model trained on a larger data to select the best DP candidate.

3. Finally, we apply the DP generator to dialogue machine translation by: 1) training an additional translation model on the DP-inserted corpus; 2) completing the input sentences with the pronouns generated from the DP generation model, before feeding them to the translation system.

[1] We extract more than 6 million quotes pairs from about 5,000 movies or TV episodes.

Experimental results on a Chinese-English dialogue translation task show that the proposed approach improves translation performance by 0.7 BLEU point using the additional translation model trained on the DP-inserted corpus. It achieves a further improvement of 1.0 BLEU point by completing the input sentences with the pronouns generated from the DP generation model.

The rest of the paper is organized as follows. In Section 2, we describe related work on DP translation. In Section 3, we present our approaches to build DP corpus, DP generator and SMT integration. Section 4 reports the experimental results of both DP generator and translation followed by a conclusion in Section 5.

2. RELATED WORK

In linguistics and translation studies, DP is regarded as implication and its explictation is observed in translations between European languages [3].

To incorporate implicit component translation in SMT, such as discourse connectives [4, 5], zero pronouns [6] and empty category [7, 8], a direct idea is to disambiguate implication categories before SMT training. Nagard and Koehn [9] present a method to aid English pronoun translation into French for SMT by integrating co-reference and unfortunately the results are not yet convincing due to the poor performance of the co-reference system. Taira et al. [10] propose both simple rule-based and manual methods to add zero pronouns at source side for Japanese–English translation. However, the BLEU scores of both systems are nearly unchanged. It indicates that 1) only considering source side and force inserting pronouns may be less principled than tackling the problem head on by integrating it into the MT itself. Hardmeier et al. [11] address the task of predicting the correct French translations of 3rd-person subject pronouns in English discourse based on neural networks. They show that its performance is competitive with that of a system with separate anaphora resolution without requiring any coreference-annotated training data. Chung and Gildea [12] systematically examine effects of EC on MT with three methods: pattern, CRF (achieves best results) and Parsing. The results show that this work can really improve the end translation even through the automatic prediction of EC is not highly accurate.

Besides, some monolingual work on co-reference, zero pronoun resolution and empty category recovering also inspire us on corpus annotation and DP prediction. Novak and Zabokrtsky [13] begin to explore cross-language differences in pronoun behavior to affect the CR results. They propose a machine learning approach to cross-lingual CR focusing on Czech pronouns. The experiment shows that bilingual feature sets are helpful to CR. Some Chinese zero pronoun resolution results are reported [14, 15, 6] based on small-scale, monolingual and general-domain training data. However, Chinese DP generation for SMT in dialogue domain is not yet addressed.

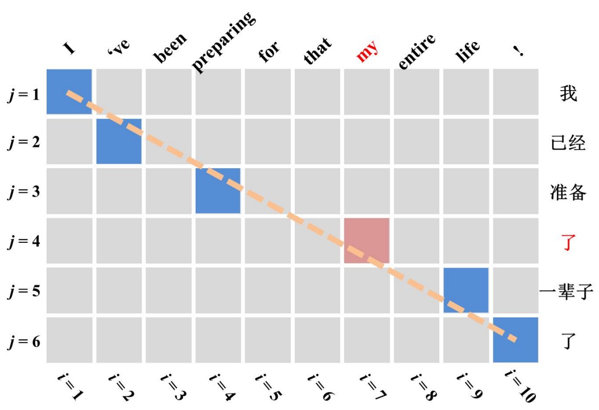

Fig. 3. Example of adding DPs using alignment results (i.e., blue blocks).

3. METHOD

The architecture of our proposed method is shown in Figure 2. It can be divided into three phases: DP training corpus annotation, DP generation and DP translation. Without loss of generality, from here on, we use Chinese as an example of pro-drop language.

3.1. DP Training Corpus Annotation

Alignments tend to occur near the diagonal, when considering the alignments as a binary $I \times J$ matrix with the cell at (i, j) be whether an alignment exists between source word i and target word j. For instance, in Figure 3, these alignment blocks follow the diagonal line. Therefore, the pronoun "*my*" ($i = 7$) in English side can be projected to an approximate area in Chinese side by considering the ahead and following alignment blocks (i.e., "*preparing*-准备" ($i = 4, j = 3$) and "*life*-一辈子" ($i = 9, j = 5$)) along the diagonal line. With this heuristic method, the red block is predicted as the potential area for the corresponding DP in Chinese side. The Chinese DP is the translation of "*my*", which is usually unique (in this case it is "我的").

However, there is still two positions to insert "DP-MY" (i.e., "我 已经 准备 DP-MY 了 一辈子 了" or "我 已经 准备 了 DP-MY 一辈子 了"). To further determine the exact position, we generate possible sentences with inserting corresponding Chinese DP into every possible position. Then we employ a large language model to score these candidates and select the one with lowest perplexity as the final result. Finally, a DP-inserted Chinese monolingual corpus is built for our DP generator training.

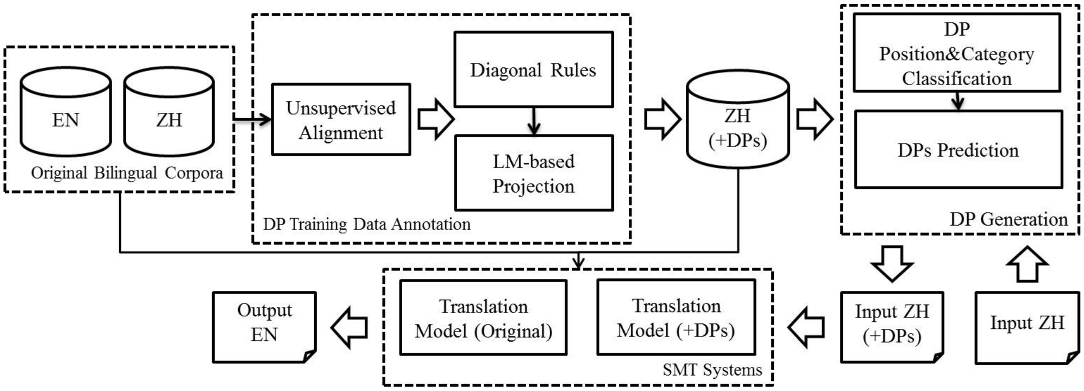

Fig. 2. Architecture of proposed method.

3.2. DP Generation

The DP generation approach is two-phase: DP position and category prediction, and DP generation.

For DP position and category classification, we regard it as a sequence labelling task. The input is the sentence consisting of a sequence of words, and the output is a sequence of tags, one for each word. In this work, we use the tag set {NULL, PE, PO, RE}, which denotes none, personal, possessive and reflexive DP.[2] In the light of the recent success of applying deep neural network technologies in NLP [16, 17], for this task we use Elman-type RNN that can be represented by

$$\mathbf{h}_t = s(\mathbf{U}\mathbf{w}_t + \mathbf{W}\mathbf{h}_{t-1}) \quad (1)$$

$$\mathbf{y}_t = g(\mathbf{V}\mathbf{h}_t) \quad (2)$$

where $s(\cdot)$ and $g(\cdot)$ are sigmoid and softmax functions respectively, and $\mathbf{U}$, $\mathbf{W}$, and $\mathbf{V}$ are the corresponding weight matrices between layers. The input vocabulary consists of words while output vocabulary consists of tags. More specifically, we follow [16] to employ a word-context window to capture short-term dependencies and use backward-forward models to capture long-term dependencies.

Once we obtain the position and category, the next step is to predict exact DP. For example, if the DP category is PO, the DP candidates are 我的 (”*my*”), 你的 (”*your*”), 我们的 (”*our*”) etc. We use an n-gram language model trained on a larger DP-inserted data to select the best DP candidate. Acutally, we tried to directly generate DPs with the RNN-based model, but obtained worse performance in a pilot study.

3.3. Integrating into Translation

The improvement of DP translation is two fold, as shown in Figure 2:

- **DP-ins. TM.** We train an additional translation model (TM) on the new parallel corpus, whose source side is inserted with DPs derived from the target side via alignment matrix (Section 3.1).
- **DP-gen. Input.** For the input sentences that have no aligned references to automatically insert DPs, we use the DP generator to generate possible DPs (Section 3.2). This makes the input sentences and *DP-ins. TM* be more consistent in terms of recalling DPs.

Beside, we build two more systems for comparison. **Baseline.** The baseline system uses the original parallel corpus and input sentences without inserting/generating DPs. **Oracle.** The oracle system uses input sentences with manually annotated DPs.

4. EXPERIMENTS

We carry out our experiments using an open-source system of the phrase-based model – Moses [18] on a Chinese-English dialogue translation tasks. Our training data contains 137,292 sentence pairs extracted from the movie subtitle website. We annotate the source side of the training data using the approach described in Section 3.1. For building the DP-inserted training data, we train a 5-gram language model on the 2.13TB Web Page Collection (SogouT) Corpus[3] using the SRI Language Toolkit [19]. To obtain a good word alignment, we run GIZA++ [20] on the training data together with another larger parallel dialogue corpus that contains 6M sentence pairs[4]. The RNN models are implemented using the common Theano neural network toolkit [21]. We extract parallel sentences from 4 movie subtitles that contains 2,601 Chinese sentences. We use 1,301 of them as the development data, and the rest of them as the test data. Note that both development and test data have only one reference. We use minimum error rate training [22] to optimize the feature

[2] If a word is tagged with PE, PO, or RE , a corresponding DP will be inserted before this word.

[3] Available at http://www.sogou.com/labs/dl/t-e.html.

[4] Extracted from Dual subtitles: http://dualsub.sourceforge.net/index.html.

DP	Data	P	R	F1
Position	Dev	0.91	0.83	0.86
+Category	Test	0.87	0.79	0.83
+Pronoun	Dev	0.80	0.77	0.78
	Test	0.76	0.72	0.74

Table 1. Evaluation of DP generation quality.

weights. For evaluation, case-insensitive NIST BLEU [23] is used to measure translation performance and F score is used to measure DP generation quality.

4.1. Evaluation of DP Generation Quality

In this section, we investigate the accuracy of the proposed approach in generating DPs. We manually annotate central pronouns on the development and test data.

We first check whether the DP annotating strategy in Section 3.1 is reasonable. To this end, we follow the strategy to automatic label the development and test data with their references. The agreements between automatic labels and manual labels are 93% and 89% on development and test data, respectively. This indicates that the automatic annotate strategy is trustworthy.

We then measure the accuracies (in terms of words) of our generation model in different phases: DP position and category prediction ("Position+Category"), and pronoun generation ("+Pronoun"). In "Position+Category", we only consider the tag for each word, while we also consider the exact pronoun for DPs. Table 1 lists the results. Remarkably, it has high F1 scores of 78% and 73% for final pronoun generation on development and test data, respectively. It demonstrates that the pronouns generated from the DP generation model are reliable to be used for end-to-end machine translation.

4.2. Evaluation of Translation Quality

Table 2 summaries the results of translation performance with different sources of DP information.

- **Baseline.** The baseline system uses the parallel corpus and input sentences without inserting/generating DPs. It achieves 19.57 and 17.66 in BLEU score on development and test data respectively. The BLEU scores are relatively low because we have only one reference.

- **+DP-ins. TM.** By using an additional translation model trained on the DP-inserted parallel corpus as described in Section 3.1, we improves the performance consistently on both development and test data. This indicates that the assumption in Section 3.1 is reasonable and the inserted DPs are helpful for machine translation.

Systems	Dev Set	Test set
Baseline	19.57	17.66
+ DP-ins. TM	20.46	18.32
+ DP-gen. Input	21.59	19.34
Oracle	24.20	20.98

Table 2. Evaluation of translation quality. "+ DP-ins. TM" denotes using an additional translation model trained on the DP-inserted training corpus, while "+ DP-gen. Input" denotes further completing the input sentences with the pronouns generated from the DP generation model.

- **+DP-gen. Input.** We can further improve translation performance by completing the input sentences with the DP generation model as described in Section 3.2. Working together with *DP-ins. TM*, we get the final improvement of 2.0 and 1.7 BLEU points on development and test data respectively. The consistence between input sentences and DP-inserted parallel corpus contributes most to the further improvement.

- **Oracle.** The oracle system uses input sentences with manually annotated DPs rather than *DP-gen. Input*. The performance gap between *oracle* and *+DP-gen. Input* shows that there is still a large space of improvement for the DP generation model.

5. CONCLUSION

Motivated by the difference in pronoun usage between Chinese and English, we investigate the translation of DP given the gold crosslingual DP annotation. Experiments show that it is crucial to identify implicit DP that require explicitation in order to improve DP translation. Analysis shows that insertion of DPs affects the decoder in a large context, which improves the overall SMT performance. The future direction of our work is on how to identify implicit DP in translation with consideration in discourse structure analysis.

Acknowledgments

This research is supported by ADAPT Centre for Digital Content Technology funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund, and also supported by the DCU-Huawei Joint Project (Grant 201504032). It is partly supported by the Open Projects Program of National Laboratory of Pattern Recognition (Grant 201407353) and the Open Projects Program of Centre of Translation of GDUFS (Grant CTS201501).

6. REFERENCES

[1] C.-T. James Huang, "On the distribution and reference of empty pronouns," *Linguistic Inquiry*, vol. 15, no. 4, pp. 531–574, 1984.

[2] Martin Haspelmath, "The european linguistic area: Standard average european," in *Language Typology and Language Universals (Volume 2)*, 2001, pp. 1492–1510.

[3] Viktor Becher, *Explicitation and implicitation in translation*, PhD thesis, 2011.

[4] Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni, "Multilingual annotation and disambiguation of discourse connectives for machine translation," in *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2011, pp. 194–203.

[5] Longyue Wang, Chris Hokamp, Tsuyoshi Okita, Xiaojun Zhang, and Qun Liu, "The dcu discourse parser for connective, argument identification and explicit sense classification," *CoNLL 2015*, p. 89, 2015.

[6] Chen Chen and Vincent Ng, "Chinese zero pronoun resolution:some recent advances," in *Proceedings of EMNLP2013*, 2013, pp. 1360–1365.

[7] Yaqin Yang and Nianwen Xue, "Chasing the ghost: recovering empty categories in the chinese treebank," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 2010, pp. 1382–1390.

[8] Mark Johnson, "A simple pattern-matching algorithm for recovering empty nodes and their antecedents," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 136–143.

[9] Ronan Le Nagard and Philipp Koehn, "Aiding pronoun translation with co-reference resolution," in *Proceedings of ACL2010*, 2010, pp. 252–261.

[10] Hirotoshi Taira, Katsuhito Sudoh, and Masaaki Nagata, "Zero pronoun resolution can improve the quality of je translation," in *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2012, pp. 111–118.

[11] Christian Hardmeier, Jorg Tiedemann, and Joakim Nivre, "Latent anaphora resolution for cross-lingual pronoun prediction," in *Proceedings of EMNLP 2013*, 2013, pp. 380–391.

[12] Tagyoung Chung and Daniel Gildea, "Effects of empty categories on machine translation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 636–645.

[13] Michal Novak and Zdenek Zabokrtsky, "Cross-lingual coreference resolution of pronouns," in *Proceedings of COLING 2014*, 2014, pp. 14–24.

[14] Shanheng Zhao and Hwee Tou Ng, "Identification and resolution of chinese zero pronouns: A machine learning approach," in *Proceedings of EMNLP 2007*, 2007, pp. 541–550.

[15] Fang Kong and Guodong Zhou, "A tree kernelbased unified framework for chinese zero anaphora resolution," in *Proceedings of EMNLP2010*, 2010, pp. 882–891.

[16] Gregoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding," in *Proceedings of INTERSPEECH 2013*, 2013.

[17] Christian Raymond and Giuseppe Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Proceedings of INTERSPEECH 2007*, 2007.

[18] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst, "Moses: open source toolkit for statistical machine translation," in *ACL 2007*, 2007.

[19] Andreas Stolcke, "Srilm - an extensible language modeling toolkit," in *Proceedings of Seventh International Conference on Spoken Language Processing*, 2002, pp. 901–904.

[20] Franz J. Och and Hermann Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[21] James Bergstra, Olivier Breuleux, Frederic Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio, "Theano: A cpu and gpu math expression compiler," in *Proceedings of Python for Scientific Computing Conference (SciPy)*, 2010, pp. 1–7.

[22] Franz Josef Och, "Minimum error rate training in statistical machine translation," in *ACL 2003*, 2003.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL 2002*, 2002.