Noname manuscript No.
(will be inserted by the editor)

A Novel and Robust Approach for Pro-Drop Language Translation

Longyue Wang · Zhaopeng Tu · Xiaojun Zhang* · Siyou Liu · Hang Li · Andy Way · Qun Liu

Received: date / Accepted: date

Abstract A significant challenge for machine translation (MT) is the phenomena of dropped pronouns (DPs), where certain classes of pronouns are frequently dropped in the source language but should be retained in the target language. In response to this common problem, we propose a semi-supervised approach with a universal framework to recall missing pronouns in translation. Firstly, we build training data for DP generation in which the DPs are automatically labelled according to the alignment information from a parallel corpus. Secondly, we build a deep learning-based DP generator for input sentences in decoding when no corresponding references exist. More specifically, the generation has two phases: (1) DP position detection, which is modeled as a sequential labelling task with recurrent neural networks; and (2) DP prediction, which employs a multilayer perceptron with rich features. Finally, we integrate the above outputs into our statistical MT (SMT) system to recall missing pronouns by both extracting rules from the DP-labelled training data and translating the DP-generated input sentences. To validate the robustness of our approach, we investigate our approach on both Chinese–English and Japanese–English corpora extracted from movie subtitles. Compared with an

* This work is most done while working in ADAPT Centre, Dublin City University.

Longyue Wang, Andy Way, Qun Liu
ADAPT Centre, School of Computing, Dublin City University, Ireland
E-mail: {longyue.wang, andy.way, qun.liu}@adaptcentre.ie

Zhaopeng Tu, Hang Li
Noah's Ark Lab, Huawei Technologies, Hong Kong
E-mail: {tu.zhaopeng, hangli.hl}@huawei.com

Xiaojun Zhang
Division of Literature and Languages, University of Stirling, UK
E-mail: xiaojun.zhang@stir.ac.uk

Siyou Liu
School of Languages and Translation, Macao Polytechnic Institute, Macau
E-mail: violetal@ipm.edu.mo

SMT baseline system, experimental results show that our approach achieves a significant improvement of +1.58 BLEU points in translation performance with 66% F-score for DP generation accuracy for Chinese–English, and nearly +1 BLEU point with 58% F-score for Japanese–English. We believe that this work could help both MT researchers and industries to boost the performance of MT systems between pro-drop and non-pro-drop languages.

Keywords pro-drop language · dropped pronoun annotation · dropped pronoun generation · machine translation · recurrent neural networks · multilayer perceptron · semi-supervised approach

1 Introduction

In pro-drop languages, certain classes of words can be omitted to make the sentence compact yet comprehensible when the identity of the pronouns can be inferred from the context. These omissions may not be problems for humans since people can easily recall the missing pronouns from the context. However, this poses difficulties for statistical machine translation (SMT) from pro-drop languages to non-pro-drop languages, since translation of such missing pronouns cannot be normally reproduced.

Among major languages, for example, Chinese and Japanese are pro-drop languages (Huang, 1984; Nakamura, 1987), while English is not (Haspelmath, 2001). Without loss of generality, we take both Chinese–English and Japanese–English examples to illustrate this phenomenon. As shown in Figure 1, *Sentences 1–2* show DP examples in Chinese–English, in which, the subject pronouns "你 (*you*)", "我 (*I*)" and the object pronouns "它 (*it*)", "你 (*you*)" are all omitted in the Chinese side. Furthermore, *Sentences 3–4* are Japanese–English examples, in which the subject pronouns "あなた (*you*)", "私 (*I*)" and the object pronouns "それ (*it*)" with their corresponding particles (e.g. "を", "は") are also omitted on the Japanese side.

We validate this finding by analysing large Chinese–English and Japanese–English corpora, which consist of sentence pairs extracted from movie and TV episode subtitles. In around 1M Chinese–English sentence pairs, we found that there are 6.5M Chinese pronouns and 9.4M English pronouns, which shows that more than 2.9 million Chinese pronouns are missing. Furthermore, in about 1.5M Japanese–English sentence pairs, there are 0.6M Japanese pronouns and 1.7M English pronouns, which shows that more than 1.1 million Japanese pronouns are missing.

To tackle the problem of omissions occurring in translation between pro-drop and non-pro-drop languages, intuitively we propose to find a general and replicable method of improving translation quality (Wang et al, 2016a,b). Becher (2011) predicted that every instance of explicitation and implicitation can be explained as a result of lexicogrammatical and/or pragmatic factors. Therefore, the task of DP translation from a pro-drop language to a non-pro-drop language should consist of making explicit what is only implied in one of the languages. Thus, the questions are (1) how to find this implicit knowledge

Fig. 1 Examples of dropped pronouns in Chinese–English (1–2) and Japanese–English (3–4) parallel corpora. The pronouns in the brackets are missing.

in the source language and (2) which DP should be generated in the target language.

The main challenge of this research is that training data for DP generation are scarce. Most current work either applies manual annotation (Yang et al, 2015) or uses existing but small-scale resources such as the Penn Treebank (Chung and Gildea, 2010; Xiang et al, 2013). In contrast, we explore an unsupervised approach to annotate DPs. Inspired by an initial idea that two languages are more informative than one (Dagan et al, 1991; Burkett et al, 2010), we propose to automatically build a large-scale training corpus for DP generation using alignment information from parallel corpora. The reason is that parallel corpora available in SMT can be used to project the missing pronouns from the target side (i.e. non-pro-drop language) to the source side (i.e. pro-drop language). To this end, we propose a simple but effective method: a bi-directional search algorithm with Language Model (LM) scoring. The LMs should be trained on large corpora in different domains from DP generation data, because the frequencies and types of DPs are very different in different domains or genres.

After building the training data for DP generation, we apply a supervised approach to build our DP generator. We divide the DP generation task into two phases: *DP detection* (from which position a pronoun is dropped), and *DP prediction* (which pronoun is dropped). Due to the powerful capacity of feature learning and representation learning, we model the DP detection problem as sequential labelling with Recurrent Neural Networks (RNNs) and model the prediction problem as classification with Multi-Layer Perceptron (MLP) using features at various levels: from lexical, through contextual, to syntax.

Finally, we integrate the DP generator into SMT system. We improve the translation of missing pronouns by explicitly recalling DPs for both parallel

data and monolingual input sentences. More specifically, we extract an additional rule table from the DP-inserted parallel corpus to produce a "pronoun-complete" translation model. In addition, we pre-process the input sentences by inserting possible DPs via the DP generation model. This makes the input sentences more consistent with the additional pronoun-complete rule table. To alleviate the propagation of DP prediction errors, we feed the translation system N-best prediction results via confusion network decoding (Rosti et al, 2007).

To validate the effect of the proposed approach, we carried out experiments on both Chinese–English (ZH–EN) and Japanese–English (JA–EN) translation tasks. Experimental results on large-scale subtitle corpora show that our approach improves translation performance by +0.61/+0.32 (ZH–EN/JA–EN) BLEU points (Papineni et al, 2002) using the additional translation model trained on the DP-inserted corpus (Koehn and Schroeder, 2007; Axelrod et al, 2011; Xu et al, 2007). Using such a model together with DP-generated input sentences achieves a further improvement. Furthermore, translation performance with N-best integration is much better than its 1-best counterpart (e.g. +0.84 and +0.84/+0.71 BLEU points on ZH–EN/JA–EN).

Generally, the contributions of this paper include the following:

- We propose an automatic method to build a large-scale DP training corpus. Given that the DPs are annotated in the parallel corpus, models trained on this data are more appropriate to the MT task;
- Benefiting from representation learning, our deep learning-based generation models are able to avoid the complex feature-engineering work while still yielding encouraging results;
- To decrease the negative effects on translation caused by inserting incorrect DPs, we force the SMT system to arbitrate between multiple ambiguous hypotheses from the DP predictions;
- We design a universal framework with these proposed pipeline components, in which each component can be evaluated and optimized in isolation;
- To demonstrate the robustness of our approaches, we evaluate our approach on both Chinese–English and Japanese–English translation tasks and compare results against a baseline SMT system.

The rest of the paper is organized as follows. Without loss of generality, we introduce the fundamental knowledge of English, Chinese and Japanese pronouns in Section 2. Section 3 is the literature review on related work. In Section 4, we describe our approaches to building the DP corpus, DP generator and SMT integration. The experimental results for both the DP generator and translation are reported in Section 5. Section 6 analyses some real examples, which is followed by our conclusion in Section 7.

2 Pronouns in English, Chinese and Japanese

In this section, we first review the characteristics of pronouns in English, Chinese and Japanese, respectively. We then discuss the differences and similari-

Category	Subject	Object	Possessive Adjective	Possessive	Reflexive
1st SG	I	me	my	mine	myself
2nd SG	you	you	your	yours	yourself
3rd SGM	he	him	his	his	himself
3rd SGF	she	her	her	hers	herself
3rd SGN	it	it	its	its	itself
1st PL	we	us	our	ours	ourselves
2nd PL	you	you	your	yours	yourselves
3rd PL	they	them	their	theirs	themselves

Table 1 English pronouns and their categories (abbreviations: person type = 1st, 2nd, 3rd, singular = SG, plural = PL, male = M, female = F and neutral = N).

Category	Subject/Object	Possessive (+ particle "的")	Reflexive (+ word "自己")
1st SG	我 (*I*/*me*)	我 的 (*my*/*mine*)	我 自己 (*myself*)
2nd SG	你 (*you*)	你 的 (*your*/*yours*)	你 自己 (*yourself*)
3rd SGM	他 (*he*/*him*)	他 的 (*his*)	他 自己 (*himself*)
3rd SGF	她 (*she*/*her*)	她 的 (*her*/*hers*)	她 自己 (*herself*)
3rd SGN	它 (*it*)	它 的 (*its*)	它 自己 (*itself*)
1st PL	我们 (*we*/*us*)	我们 的 (*our*/*ours*)	我们 自己 (*ourselves*)
2nd PL	你们 (*you*)	你们 的 (*your*/*yours*)	你们 自己 (*yourselves*)
3rd PLM	他们 (*they*/*them*)	他们 的 (*their*/*theirs*)	他们 自己 (*themselves*)
3rd PLF	她们 (*they*/*them*)	她们 的 (*their*/*theirs*)	她们 自己 (*themselves*)
3rd PLN	它们 (*they*/*them*)	它们 的 (*their*/*theirs*)	它们 自己 (*themselves*)

Table 2 Correspondence of pronouns in Chinese–English (use the same abbreviations in Table 1).

ties in Chinese–English and Japanese–English language pairs from a bilingual point of view.

In English, Quirk et al (1985) classifies the principal pronouns into three groups: personal pronouns, possessive pronouns and reflexive pronouns, defining them as central pronouns. As shown in Table 1, all of the central pronouns have diverse forms to demonstrate or indicate different person, number, gender and function. For example, the pronoun "*we*" represents the first person in plural form and functions as subject in a sentence, while another pronoun "*him*" indicates the masculine third person in singular form and functions as a object of a verb.

Generally, Chinese pronouns correspond to the personal pronouns in English, and the Chinese pronominal system is relatively simple as there is no inflection, conjugation, or case makers (Li and Thompson, 1989). Thus, there is no difference between subjective and objective pronouns (we call them "basic pronouns"). Besides, possessive and reflexive pronouns can be generated by adding some particle or modifier based on the basic pronouns. We show the Chinese–English pronouns in Table 2.

Category	Japanese Pronouns (Subject/Object)
1st SG	私, 我, 俺, 僕, 儂, 家, etc. (*I*/*me*)
2nd SG	お前, おまえ, なん, 君, 貴方, あなた, あんた, 貴様, etc. (*you*)
3rd SGM	そいつ, あいつ,あの人, あの方, 彼, etc. (*he/him*)
3rd SGF	そいつ, あいつ, あの人, あの方, 彼女, etc. (*she/her*)
3rd SGN	そいつ. (*it*)
1st PL	我々, 我等, etc. (*we*/*us*)
2nd PL	お前, おまえ, なん, 君, 貴方, あなた, あんた, 貴様, etc. (*you*)
3rd PL	彼等 (*they*/*them*)

Table 3 Correspondence of pronouns in Japanese–English (use the same abbreviations in Table 1).

As shown in Table 2, the Chinese pronouns are not strictly consistent to the English pronouns. On the one hand, one Chinese pronoun can be translated to several English pronouns (one-to-many). For instance, the Chinese pronoun "我" can be mapped to both the subjective personal pronoun "*I*" and the objective personal pronoun "*me*". On the other hand, there are also some many-to-one cases. For example, the pronouns "他们", "她们", "它们" can all be translated into the English pronoun "*they*", because the Chinese pronominal system considers gender for third person plural pronouns while English does not. "你们/你 - *you*" is another many-to-one case, because the English pronominal system does not differentiate between the singular and plural forms for second person pronoun while the Chinese system does.

Similar to Chinese, the Japanese pronouns can be altered to possessive and reflexive through adding the particle "の" or modifier "自分" to the basic pronouns, respectively. Besides, the same form of pronouns in Japanese can be used to function as subject or object with different particles. For example, the particle "は" comes after the subjective pronouns, while the particle "を" occurs after the objective pronouns.

In Table 3, we only list the most commonly used forms of subjective/objective pronouns, because possessive and reflexive pronouns can be generated by adding corresponding particles. Different from English and Chinese, Japanese has a large number of pronoun variations. The Japanese pronominal system considers more factors such as gender, age, and relative social status of the speaker and audience. For instance, the first person singular pronoun "私" is used in formal situations, while "僕" and "俺" refer to male pronouns and are normally used in informal contexts. Besides, "儂" is mostly used in old Japanese society or to indicate old male characters, while "家" is frequently used by young girls.

3 Related Work

Natural language tasks in one language can be improved by exploiting translations in another language. This observation has formed the basis for impor-

tant work on syntax projection across languages (Yarowsky and Ngai, 2001; Hwa et al, 2005; Kuzman Ganchev and Taskar, 2009) and unsupervised syntax induction in multiple languages (Snyder et al, 2009), as well as other tasks, such as cross–lingual named entity recognition (Huang and Vogel, 2002; Moore, 2003; Wang and Manning, 2014) and information retrieval (Si and Callan, 2005). In all of these cases, multilingual models yield increased accuracy because different languages present different ambiguities and therefore offer complementary constraints on the shared underlying labels.

There is some work related to DP generation. One is zero pronoun resolution (ZP), which is a sub-direction of co-reference resolution (CR). The difference to our task is that ZP contains three steps (namely ZP detection, anaphoricity determination and co-reference link) whereas DP generation only contains the first two steps. Some researchers (Zhao and Ng, 2007; Kong and Zhou, 2010; Chen and Ng, 2013) propose rich features based on different machine-learning methods. For example, Chen and Ng (2013) propose an SVM classifier using 32 features including lexical, syntax and grammatical roles etc., which are very useful in the ZP task. However, most of their experiments are conducted on a small-scale corpus (i.e. OntoNotes)[1] and performance drops correspondingly when using a system-parse tree compared to the gold standard one. Novak and Zabokrtsky (2014) explore cross-language differences in pronoun behavior to affect the CR results. The experiment shows that bilingual feature sets are helpful to CR. Another line related to DP generation is using a wider range of empty categories (EC) (Yang and Xue, 2010; Cai et al, 2011; Xue and Yang, 2013), which aims to recover long-distance dependencies, discontinuous constituents and certain dropped elements[2] in phrase structure treebanks (Xue et al, 2005). This work mainly focuses on sentence-internal characteristics as opposed to contextual information at the discourse level. More recently, Yang et al (2015) explored DP recovery for Chinese text messages based on both lines of work.

The above methods can also be used for DP translation using SMT (Chung and Gildea, 2010; Le Nagard and Koehn, 2010; Taira et al, 2012; Xiang et al, 2013). Taira et al (2012) propose both simple rule-based and manual methods to add zero pronouns on the source side for Japanese–English translation. However, the BLEU scores of both systems are nearly identical, which indicates that only considering the source side and forcing the insertion of pronouns may be less principled than tackling the problem head on by integrating them into the SMT system itself. Le Nagard and Koehn (2010) present a method to aid English pronoun translation into French for SMT by integrating CR. Unfortunately, their results are not convincing due to the poor performance of the CR method (Pradhan et al, 2012). Chung and Gildea (2010) systematically examine the effects of EC on MT with three methods: pattern, CRF (which achieves best results) and parsing. The results show that this work can really

[1] It contains 144K coreference instances, but only 15% of them are dropped subjects.

[2] EC includes trace markers, dropped pronoun, big PRO etc, while we focus only on dropped pronoun.

improve the end translation even though the automatic prediction of EC is not highly accurate.

4 Methodology

We propose a universal architecture for our method as shown in Figure 2, which can be divided into three main components: DP training data annotation, DP generation, and SMT integration. Given a parallel corpus, we automatically annotate with DPs by projecting aligned pronouns from the target side to the source side. With the annotated DP training corpus, we then propose a supervised approach to DP generation. Finally, we integrate the DP generator into MT using various methods. In this work, we mainly focus on subjective, objective and possessive pronouns (as described in Section 2) without considering reflexive ones, because of the low frequency of reflexive pronouns in our corpora. To make the Japanese pronouns simple, we replace all pronoun variations with a unified one in our corpora.

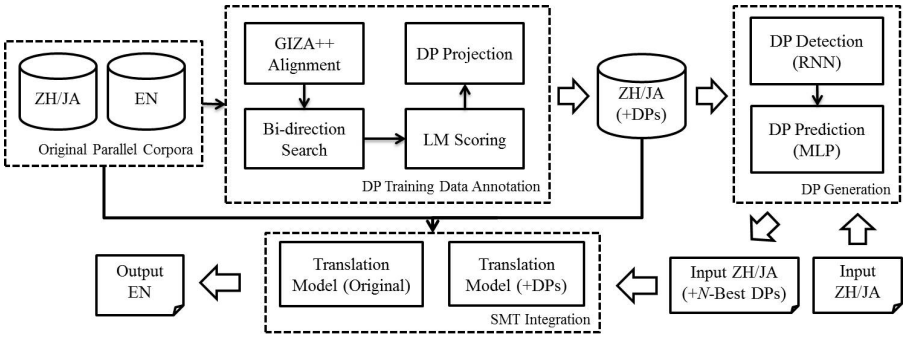

Fig. 2 Architecture of our proposed method.

Algorithm 1 Bidirectional search algorithm in MATLAB^{TM}

```
function [DP_start, DP_end] = BidirectionalSearch(Matrix, Misalign)
    row = sum(Matrix, 1);
    row_true = find(row == 1);
    left_side = row_true(row_true < Misalign);
    DP_start = find(Matrix(:, left_side(end)) == 1);
    right_side = row_true(row_true > Misalign);
    DP_end = find(Matrix(:, right_side(1)) == 1);
end
```

4.1 DP Training Corpus Annotation

We propose an approach to automatically annotate DPs by utilizing alignment information. Given a parallel corpus, we first use an unsupervised word alignment method (Och and Ney, 2003; Tu et al, 2012) to produce a word alignment. From observing the alignment matrix, we found it is possible to detect DPs by projecting misaligned pronouns from the non-pro-drop target side (e.g. English) to the pro-drop source side (e.g. Chinese).

Therefore, we propose a bidirectional search algorithm as shown in Algorithm 1. Given the alignment matrix $Matrix$ and the misaligned pronoun position $Misalign$, the algorithm searches from $Misalign$ to the beginning and the end of the target sentence, respectively. If one word in the target language is aligned with one word in the source language, we call them aligned words (the value is set as 1), otherwise they are considered to be misaligned words (the value is set as 0). The algorithm tries to find the nearest preceding and following aligned words around $Misalign$, and then to project them to the DP positions (*start* or *end*) on the source side.

As shown in Figure 3, we use a Chinese–English example to illustrate our idea. We consider the alignments as a binary $I \times J$ matrix with the cell at (i, j), to decide whether an alignment exists between Chinese word i and English word j. For each pronoun on the English side (e.g. "*I*", "*my*"), we first check whether it has an aligned pronoun on the Chinese side. We find that the pronoun "*my*" ($i = 7$) is not aligned to any Chinese word and possibly corresponds to a DP_{MY}. To determine the possible positions of DP_{MY} on the Chinese side, we employ a diagonal heuristic based on the observation that there exists a diagonal rule in the local area of the alignment matrix. With this heuristic method, the DP_{MY} can be projected to an approximate area (red block) on the Chinese side by considering the preceding 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.

However, there are still two possible positions to insert DP_{MY} (i.e. the two gaps before or after the word "了"). To further determine the exact position of DP_{MY}, we generate possible sentences by inserting the corresponding Chinese translation of DP into every possible position (i.e., "我 已经 准备 **我的** 了 一辈子 了" or "我 已经 准备 了 **我的** 一辈子 了"). The Chinese translation of DP can be determined by using its English pronouns according to Table 2. Note that some English pronouns may correspond to more than one Chinese pronoun, such as "*they* - 他们 / 她们 / 它们". In this case, we use all the corresponding Chinese pronouns as the candidates. Then we employ an n-gram language model (LM) to score these candidates and select the one with the lowest perplexity as the final result. This LM-based projection is based on the observation that the amount and type of DPs are very different in different genres. We hypothesize that the DP position can be determined by utilizing the inconsistency of DPs in different domains. Therefore, the LM is trained on a large amount of Chinese news data or Japanese combined domain of data (detailed in Section 5). In order to reduce the problem of incorrect DP

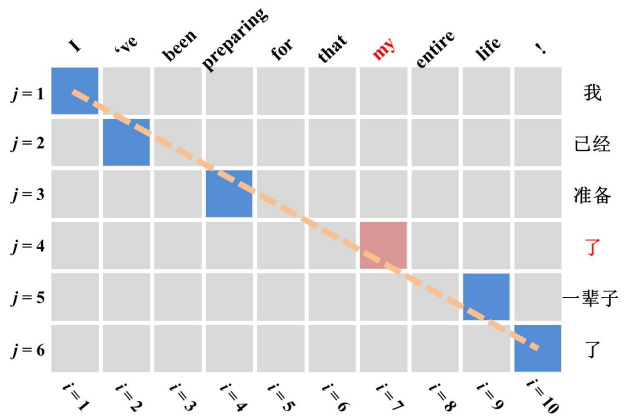

Fig. 3 Example of DP projection using alignment results (i.e. blue blocks).

insertion caused by incorrect alignment, we use a large amount of additional parallel corpus to improve the quality of alignment. Finally, a DP-inserted Chinese monolingual corpus is built for our DP generator training.

4.2 DP Generation

In light of the recent success of applying deep neural network technologies in natural language processing (Raymond and Riccardi, 2007; Mesnil et al, 2013), we propose a neural network-based DP generator via the DP-inserted corpus. We first employ an RNN to predict the DP position, and then train a classifier using multilayer perceptrons to generate the DP results.

4.2.1 DP detection

The task of DP position detection is to label words if there are pronouns missing before the words, which can intuitively be regarded as a sequence labelling problem. We expect the output to be a sequence of labels $y^{(1:n)} = (y^{(1)}, y^{(2)}, \cdots, y^{(t)}, \cdots, y^{(n)})$ given a sentence consisting of words $w^{(1:n)} = (w^{(1)}, w^{(2)}, \cdots, w^{(t)}, \cdots, w^{(n)})$, where $y^{(t)}$ is the label of word $w^{(t)}$. In our task, there are two labels $L = \{NA, DP\}$ (corresponding to non-pro-drop or pro-drop pronouns), thus $y^{(t)} \in L$.

Word embeddings (Mikolov et al, 2013) are used for our generation models: given a word $w^{(t)}$, we try to produce an embedding representation $\mathbf{v}^{(t)} \in \mathbb{R}^d$ where d is the dimension of the representation vectors. In order to capture short-term temporal dependencies, we feed the RNN unit a window of context, as in Equation (1):

$$\mathbf{x_d}^{(t)} = \mathbf{v}^{(t-k)} \oplus \cdots \oplus \mathbf{v}^{(t)} \oplus \cdots \oplus \mathbf{v}^{(t+k)} \tag{1}$$

where k is the window size.

We employ an RNN (Mesnil et al, 2013) to learn the dependency of sentences, which can be formulated as Equation (2):

$$\mathbf{h}^{(t)} = f(\mathbf{U}\mathbf{x_d}^{(t)} + \mathbf{V}\mathbf{h}^{(t-1)}) \tag{2}$$

where $f(x)$ is a sigmoid function at the hidden layer. $\mathbf{U}$ is the weight matrix between the raw input and the hidden nodes, and $\mathbf{V}$ is the weight matrix between the context nodes and the hidden nodes. At the output layer, a softmax function is adopted for labelling, as in Equation (3):

$$y^{(t)} = g(\mathbf{W_d}\mathbf{h}^{(t)}) \tag{3}$$

where $g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}$, and $\mathbf{W_d}$ is the output weight matrix.

4.2.2 DP prediction

Once the DP position is detected, the next step is to determine which pronoun should be inserted based on this result. Accordingly, we train a m-class classifier ($m = 20$ in our experiments), where each class refers to a distinct Chinese/Japanese pronoun category in Section 2.

We select a number of features based on previous work (Xiang et al, 2013; Yang et al, 2015), including lexical, contextual, and syntax features (as shown in Table 4). We set p as the DP position, S as the window size surrounding p, and X, Y as the window size surrounding current sentence (the one contains p). For Features 1–4, we extract words, POS tags and pronouns around p. For Features 5–8, we also consider the pronouns and nouns in X / Y preceding or following sentences. For Features 9–10, in order to model the syntactic relation, we use a path feature, which is the combined tags of the sub-tree nodes from p / $(p-1)$ to the root. Note that Features 3–6 only consider pronouns that were not dropped. Each unique feature is treated as a word, and assigned a "word embedding". The embeddings of the features are then fed to the neural network. We fix the number of features for the variable-length features, where missing ones are tagged as *None*. Accordingly, all training instances share the same feature length. For the training data, we sample all DP instances from the corpus (annotated by the method in Section 4.1). During decoding, p can be given by our DP detection model.

We employ a feed-forward neural network with four layers. The input $\mathbf{x_p}$ comprises the embeddings of the set of all possible feature indicator names. The middle two layers $\mathbf{a}^{(1)}$, $\mathbf{a}^{(2)}$ use Rectified Linear function R as the activation function, as in Equation (4)–(5):

$$\mathbf{a}^{(1)} = R(\mathbf{b}^{(1)} + \mathbf{W_p}^{(1)}\mathbf{x_p}) \tag{4}$$

$$\mathbf{a}^{(2)} = R(\mathbf{b}^{(2)} + \mathbf{W_p}^{(2)}\mathbf{a}^{(1)}) \tag{5}$$

where $\mathbf{W_p}^{(1)}$ and $\mathbf{b}^{(1)}$ are the weights and bias connecting the first hidden layer to second hidden layer; and so on. The last layer $\mathbf{y_p}$ adopts the softmax

Feature Set	ID.	Description
Lexical	1	S surrounding words around p
	2	S surrounding POS tags around p
	3	preceding pronoun in the same sentence
	4	following pronoun in the same sentence
Context	5	pronouns in preceding X sentences
	6	pronouns in following X sentences
	7	nouns in preceding Y sentences
	8	nouns in following Y sentences
Syntax	9	path from current word (p) to the root
	10	path from preceding word ($p-1$) to the root

Table 4 List of features.

function g, as in Equation (6):

$$\mathbf{y_p} = g(\mathbf{W_p}^{(3)}\mathbf{a}^{(2)}) \tag{6}$$

4.3 Integration into Translation

Different from the baseline SMT system that uses the parallel corpus and input sentences without inserting/generating DPs, the integration into SMT system is three fold: DP-inserted translation model (*DP-ins. TM*), DP-generated input (*DP-gen. Input*) and N-best inputs.

4.3.1 DP-inserted TM

We train an additional translation model on the new parallel corpus, whose source side is inserted with DPs derived from the target side via the alignment matrix (detailed in Section 4.1). We hypothesize that DP insertion can help to obtain a better alignment, which can benefit translation. Then the whole translation process is based on the boosted translation model, i.e. with DPs inserted. As far as TM combination is concerned, we directly feed Moses the multiple phrase tables. The gain from the additional TM is mainly from complementary information about the recalled DPs from the annotated data.

4.3.2 DP-generated input

Another option is to pre-process the input sentence by inserting possible DPs with the DP generation model (detailed in Section 4.2) so that the DP-inserted input (Input ZH+DPs) is translated. The predicted DPs would be explicitly translated into the target language, so that the possibly missing pronouns in the translation might be recalled. This makes the input sentences and DP-inserted TM more consistent in terms of recalling DPs.

4.3.3 N-best inputs

However, the above method suffers from a major drawback: it only uses the 1-best prediction result for decoding, which potentially introduces translation mistakes due to the propagation of prediction errors. To alleviate this problem, an obvious solution is to offer more alternatives. Recent studies have shown that SMT systems can benefit from widening the annotation pipeline (Liu et al, 2009; Tu et al, 2010, 2011; Liu et al, 2013). In the same direction, we propose to feed the decoder N-best prediction results, which allows the system to arbitrate between multiple ambiguous hypotheses from upstream processing so that the best translation can be produced. The general method is to make the input with N-best DPs into a confusion network. In our experiment, each prediction result in the N-best list is assigned a weight of $1/N$.

5 Experiments

5.1 Setup

For Chinese–English training data, we extract around 1M sentence pairs (movie or TV episode subtitles) from two subtitle websites (Wang et al, 2016c).[3] For Japanese–English training data, we use OpenSubtitles2016 corpus[4]. We manually create both development and test sets with DP annotation. The detailed statistics of data are listed in Table 5 and 6. Note that all sentences maintain their contextual information at the discourse level, which can be used for feature extraction in Section 4.2. There are two different language models for the DP annotation (detailed in Section 4.1) and translation tasks (detailed in Section 4.3), respectively: one is trained on the Chinese News Collection Corpus[5] or use Japanese combined corpus[6] while the other one is trained on all extracted 7M English subtitle data.

We carry out our experiments using the phrase-based SMT model in Moses (Koehn et al, 2007) on a Chinese–English and Japanese–English translation task. Furthermore, we train 5-gram language models using the SRI Language Toolkit (Stolcke, 2002). To obtain a good word alignment, we run GIZA++ (Och and Ney, 2003) on the training data together with another larger parallel subtitle corpora.[7] As our annotation method (Section 4.1) relies on the quality

[3] Avaliable at `http://www.opensubtitles.org` and `http://weisheshou.com`.

[4] We use part of Japanese–English data, which is available at `http://opus.lingfil.uu.se/OpenSubtitles2016.php`.

[5] Available at `http://www.sogou.com/labs/dl/ca.html`.

[6] We collect a number of monolingual corpora such as KFTT (`http://www.phontron.com/kftt`), NTCIR (`http://warehouse.ntcir.nii.ac.jp/openaccess/rite/10RITE-Japanese-wiki.html`) and Wikipedia XML Corpus (`http://www-connex.lip6.fr/~denoyer/wikipediaXML`).

[7] Our Chinese–English additional corpus contains more than 9M sentence pairs (Zhang et al, 2014) and the Japanese–English additional corpus contains 1.5M sentence pairs (Lison and Tiedemann, 2016).

Corpus	Lang.	Sentences	Pronouns	Ave. Len.
Train	ZH	1,037,292	604,896	5.91
	EN	1,037,292	816,610	7.87
Dev	ZH	1,086	756	6.13
	EN	1,086	1,025	8.46
Test	ZH	1,154	762	5.81
	EN	1,154	958	8.17

Table 5 Statistics of Chinese–English corpora.

Corpus	Lang.	Sentences	Pronouns	Ave. Len.
Train	JA	501,119	178,823	8.55
	EN	501,119	554,561	8.65
Dev	JA	1,146	413	8.24
	EN	1,146	1,274	8.84
Test	JA	1,150	427	8.11
	EN	1,150	1,280	8.17

Table 6 Statistics of Japanese–English corpora.

of alignment, we employ "intersection" alignment method, which has high precision, but low recall. We use minimum error rate training (Och, 2003) to optimize the feature weights.

The RNN models are implemented using the common Theano neural network toolkit (Bergstra et al, 2010). We use a pre-trained word embedding via a lookup table. We use the following settings: windows = 5, the size of the single hidden layer = 200, iterations = 10, embeddings = 200. The MLP classifier uses random initialized embeddings, with the following settings: the size of the single hidden layer = 200, embeddings = 100, iterations = 200.

For end-to-end evaluation, case-insensitive BLEU (Papineni et al, 2002) is used to measure translation performance and micro-averaged F-score is used to measure DP generation quality.

5.2 Evaluation of DP Generation

We first check whether our DP annotation strategy is reasonable. To this end, we follow the strategy to automatically and manually label the source sides of the development and test data with their target sides. The results are shown in Table 7. For Chinese–English, the agreement between automatic labels and manual labels on DP prediction are 94% and 95% on development and test data and on DP generation are 92% and 92%, respectively. However, the agreements of Japanese–English sets are lower. The main reason is that Japanese is a subject—object—verb (SOV) language while Chinese and English are subject–verb-–object (SVO) languages. The difference of language ordering between Japanese and English make the bidirectional search algorithm more difficult to map. Generally, these results (above 80%) indicate that the automatic annotation strategy is relatively trustworthy.

Language	DP Detection		DP Prediction	
	dev set	test set	dev set	test set
ZH-EN	0.94	0.95	0.92	0.92
JA-EN	0.91	0.90	0.85	0.83

Table 7 Evaluation of DP annotation quality.

We then measure the accuracy (in terms of words) of our generation models in two phases. "DP Detection" shows the performance of our sequence-labelling model based on RNN. We only consider the tag for each word (pro-drop or not pro-drop before the current word), without considering the exact pronoun for DPs. "DP Prediction" shows the performance of the MLP classifier in determining the exact DP based on detection. Thus, we consider both the detected and predicted pronouns. Table 8 lists the results of the above DP generation approaches. For Chinese, the F1 score of "DP Detection" achieves 88% and 86% on the Dev and Test set, respectively. However, it has lower F1 scores of 66% and 65% for the final pronoun generation ("DP Prediction") on the development and test data, respectively. This indicates that generating the exact DP in Chinese is a difficult task. As far as the Japanese results are concerned, the performance of DP detection and prediction is lower than Chinese. "DP Detection" achieves 81% and 80% F1 scores on the Dev and Test set, respectively, while "DP Predition" obtains 59% and 58%, respectively. Even though the DP prediction is not highly accurate, we still hypothesize that the DP generation models are reliable enough to be used for end-to-end MT. Note that we only show the results of 1-best DP generation here, but in the translation task itself, we use N-best generation candidates to recall more DPs.

Language	Set	DP Detection			DP Prediction		
		P	R	F1	P	R	F1
ZH	Dev	0.88	0.84	0.86	0.67	0.63	0.65
	Test	0.88	0.87	0.88	0.67	0.65	0.66
JA	Dev	0.83	0.80	0.81	0.61	0.58	0.59
	Test	0.81	0.79	0.80	0.60	0.57	0.58

Table 8 Evaluation of DP generation quality.

5.3 Evaluation of DP Translation

In this section, we evaluate the end-to-end translation quality by integrating the DP generation results (Section 4.3). Table 9 and 10 summarise the results of translation performance with different sources of DP information for Chinese–English and Japanese–English, respectively. "Baseline" uses the original input to feed the SMT system. "+DP-ins. TM" denotes using an additional translation model trained on the DP-inserted training corpus, while

Systems	Dev Set	Test set
Baseline	20.06	18.76
+DP-ins. TM	20.32 (+0.26)	19.37 (+0.61)
+DP-gen. Input		
1-best	20.49 (+0.43)	19.50 (+0.74)
2-best	20.15 (+0.09)	18.89 (+0.13)
4-best	20.64 (+0.58)	19.68 (+0.92)
6-best	21.61 (+1.55)	20.34 (+1.58)
8-best	20.94 (+0.88)	19.83 (+1.07)
Manual Oracle	24.27 (+4.21)	22.98 (+4.22)
Auto Oracle	23.10 (+3.04)	21.93 (+3.17)

Table 9 Evaluation of Chinese–English DP translation quality.

Systems	Dev Set	Test set
Baseline	18.24	16.54
+DP-ins. TM	18.58 (+0.34)	16.86 (+0.32)
+DP-gen. Input		
1-best	18.54 (+0.30)	16.79 (+0.25)
2-best	18.79 (+0.55)	17.08 (+0.54)
4-best	19.32 (+1.08)	17.50 (+0.96)
6-best	19.11 (+0.87)	17.41 (+0.87)
8-best	18.84 (+0.60)	17.11 (+0.57)
Manual Oracle	20.78 (+2.54)	18.84 (+2.30)
Auto Oracle	20.06 (+1.82)	18.31 (+1.77)

Table 10 Evaluation of Japanese–English DP translation quality.

"+DP-gen. Input N" denotes further completing the input sentences with the N-best pronouns generated from the DP generation model. "Oracle" uses the input with manual ("Manual") or automatic ("Auto") insertion of DPs by considering the target set. Taking "Auto Oracle" for example, we annotate the DPs via alignment information (supposing the reference is available) using the technique described in Section 4.1.

The baseline system uses the parallel corpus and input sentences without inserting/generating DPs. The Chinese–English system achieves 20.06 and 18.76 in BLEU score on the development and test data, respectively. The BLEU scores are relatively low because 1) we have only one reference, and 2) dialogue machine translation is still a challenge for the current SMT approaches. Besides, the Japanese–English system achieves 18.24 and 16.54 in BLEU score on the development and test data, respectively. Apart from the above two reasons, the BLEU scores are lower because the size of Japanese–English parallel corpus is smaller.

By using an additional translation model trained on the DP-inserted parallel corpus as described in Section 4.1, we improve the performance consistently on both development (ZH-EN: +0.26 and JA-EN: +0.34) and test data (ZH-EN: +0.61 and JA-EN: +0.32). This indicates that the inserted DPs are really helpful for SMT. Thus, the gain in the "+DP-ins TM" is mainly from the improved alignment quality.

We can further improve translation performance by completing the input sentences with our DP generation model as described in Section 4.2. We test N-best DP insertion to examine the performance, where $N = \{1, 2, 4, 6, 8\}$. For Chinese–English, working together with "DP-ins. TM", 1-best generated input already achieves +0.43 and +0.74 BLEU score improvements on development and test set, respectively. The consistency between the input sentences and the DP-inserted parallel corpus contributes most to these further improvements. As N increases, the BLEU score grows, peaking at 21.61 and 20.34 BLEU points when N=6. Thus, we achieve a final improvement of +1.55 and +1.58 BLEU points on the development and test data, respectively. However, when adding more DP candidates, the BLEU score decreases by 0.97 and 0.51. The reason for this may be that more DP candidates add more noise, which harms the translation quality. It is similar to Japanese–English results, but the improvements are relatively lower. For example, the best BLEU scores are 19.32 (+1.08) and 17.50 (+0.96) on development and test set when N=4. It shows that Japanese–English is more difficult to deal with pronoun translation problems than Chinese–English.

The oracle system uses the 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 for further improvement for the DP generation model, especially for Chinese–English.

6 Analysis and Discussion

In this section, we first select sample sentences to further investigate the effect of DP generation on translation. As Chinese–English and Japanese–English outputs have similar characteristics, we mainly take Chinese–English examples for analysis. Furthermore, we also show alignment examples to discuss Japanese–English results.

In the following sentences, we show a positive case (Case A), a negative case (Case B) and a neutral case (Case C) of translation by using DP insertion (i.e. "+DP-gen. Input 1-best" detailed in Section 4.3.2) as well as N-best case (Case D) (i.e. "+DP-gen. Input N-best" detailed in Section 4.3.3). In Cases A-C, we give (a) the original Chinese sentence and its translation generated by the baseline system, (b) the DP-inserted Chinese sentence and its translation generated by "+DP-gen. Input 1-best" system, and (c) the reference English sentence. In Case D, (a) is the original Chinese sentence and its translation, and (b)-(d) are N-best DP-generated Chinese sentences and their MT outputs, and (e) is the reference.

In Case A, the output of (a) (generated by the original Chinese sentence) is incomplete because it is missing a subject on the English side. However, by adding a DP "你 (you)" via our DP generator, "*Do you*" is produced in the output of (b). It not only gives a better translation than (a), but also makes the output a formal general question sentence. We found that inserting DPs

Case A

(a) 想不想 听 一件 奇怪的 事 ？
Wanna hear something weird ?

(b) ⟨**你**⟩ 想不想 听 一件 奇怪的 事 ？
Do ⟨**you**⟩ want to hear something weird ?

(c) Do you want to hear something weird ?

Case B

(a) 你 肯定 看过 那 电视剧 。
You must have seen that show .

(b) 你 肯定 ⟨**我**⟩ 看过 那 电视剧 。
You are sure ⟨**I**⟩ 've seen that show .

(c) You must have seen that show .

Case C

(a) 不要 告诉 瑞秋 ， 待会 见 。
Do not tell Rachel . see you later .

(b) 不要 告诉 瑞秋 ， ⟨**你**⟩ 待会 见 。
Do not tell Rachel . see ⟨**you**⟩ later .

(c) Do not tell Rachel . see you later .

into interrogative sentences helps both reordering and grammar. Generally, Case A shows that 1-best DP generation can really help translation.

In Case B, however, our DP generator mistakenly regards the simple sentence as a compound sentence and inserts the wrong pronoun "我 (I)" in (b), which causes an incorrect translation output (worse than (a)). This indicates that we need a highly accurate source-sentence parse tree for more correct detection of the antecedent of DPs. Besides, some errors are caused by pre-processing such as Chinese segmentation and part-of-speech (POS) tagging. For instance, a well-tagged sentence should be "他/PN 好/VA 有/VE 魅力/NN (He has a good charm)". However, in our experiments, the sentence is incorrectly tagged as "他/PN 好/VA 有魅力/VE" and the DP generator inserts a DP "我 (I)" between "好" and "有魅力". Therefore, our features should be extracted based on a natural language processing toolkit with good performance.

Case D

(a) 都 不会 想 我 吗 ?

Won 't even miss me ?

(b) ⟨**我**⟩ 都 不会 想 我 吗 ?

⟨**I**⟩ won 't even miss me ?

(c) ⟨**我/你 ...**⟩ 都 不会 想 我 吗 ?

⟨**You**⟩ won 't even miss me ?

(d) ⟨**我/你/他 ...**⟩ 都 不会 想 我 吗 ?

⟨**He**⟩ won 't even miss me ?

(e) You won 't even miss me ?

In Case C, the translation results are the same in (a) and (b). Such unchanged cases often occur in "fixed" linguistic chunks such as preposition phrases ("on *my* way"), greetings ("see *you* later" , "thank *you*") and interjections ("*my* God"). However, the alignment of (b) is better than that of (a) in this case. It also shows that even though the DP is inserted in a wrong place, it can still be reordered into the correct translation due to the powerful target LM. This explains why end-to-end performance can be improved even with a sub-optimal DP generator.

In Case D, (a) is the original Chinese sentence and its translation; (b) is the 1-best DP-generated Chinese sentence and its MT output; (c) stands for 2-best, 4-best and 6-best DP-generated Chinese sentences and their MT outputs (which are all the same); (d) is the 8-best DP-generated Chinese sentence and its MT output; (e) is the reference. The N-best DP candidate list is "我 (I)", "你 (You)", "他 (He)", "我们 (We)", "他们 (They)", "你们 (You)", "它 (It)" and "她 (She)". In (b), when integrating an incorrect 1-best DP into MT, we obtain the wrong translation. When considering more DPs (2-/4-/6-best) in (c), the SMT system generates a correct translation by weighting the DP candidates during decoding. When further increasing N (8-best), (d) shows a wrong translation again due to increased noise.

The Japanese–English translation is more difficult due to the different sentence structures between them. Besides, the alignment results sometimes do not follow the diagonal rules (as discussed in Section 4.1). Considering the examples in Figure 4, the left alignment box shows a simple case where the alignments follow a diagonal line. However, the right one is more complex, in which the English pronoun "me" can be projected according to local diagonal heuristics while the pronoun "You" is difficult to be projected into the correct position. Thus, the search spaces of the misaligned "You" are all the

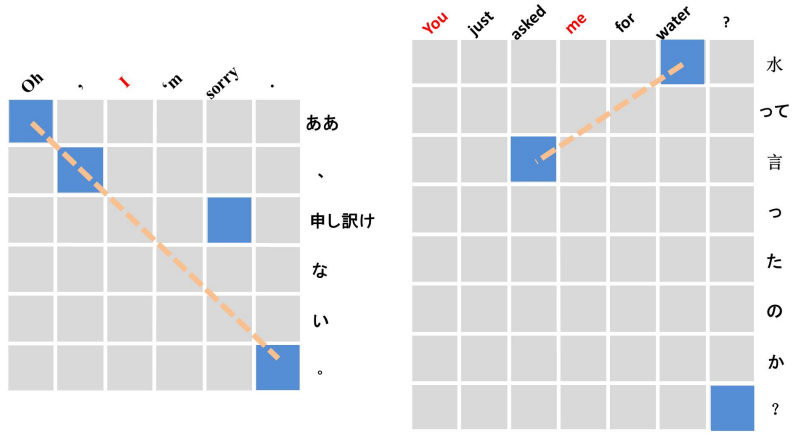

Fig. 4 Alignment results from Japanese–English corpus.

positions of the Japanese sentence with high error rate. That is why the the DP annotation quality is much lower (as shown in Table 8) than Chinese one. Furthermore, these annotation errors are propagated to the following components of the architecture (as shown in Figure 2) and harm the translation to some extent.

7 Conclusion and Future Work

In this paper, we have presented a novel approach to recall missing pronouns for machine translation from a pro-drop language to a non-pro-drop language. We first propose an automatic approach to DP annotation, which utilizes alignment matrix from parallel data and shows high consistency compared with manual annotation method. We then applied neural networks to DP detection and prediction tasks with rich features. About integration into translation, we employ confusion networks decoding with N-best DP prediction results instead of ponderously inserting only 1-best DP into input sentences. Finally we implemented above models into a well designed DP translation architecture.

Experiments on both Chinese–English and Japanese-English translation tasks show that it is crucial to identify the DP to improve the overall translation performance. Our analysis shows that insertion of DPs affects the translation to a large extent.

Our main findings in this paper are threefold:

- Bilingual information can help to build monolingual models without any manually annotated training data;
- Benefiting from representation learning, neural network-based models work well without complex feature engineering work;
- N-best DP integration works better than 1-best DP insertion;

– Our approach is robust and can be applied on pro-drop languages especially for Chinese.

In future work, we plan to extend our work to different genres, integration with neural translation system and other kinds of dropped words to validate the robustness of our approach.

Acknowledgments

This work is supported by the Science Foundation of Ireland (SFI) ADAPT project (Grant No.:13/RC/2106), and partly supported by the DCU-Huawei Joint Project (Grant No.:201504032-A (DCU), YB2015090061 (Huawei)).

References

Axelrod A, He X, Gao J (2011) Domain adaptation via pseudo in-domain data selection. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, pp 355–362

Becher V (2011) Explicitation and implicitation in translation. PhD thesis, Universität Hamburg

Bergstra J, Breuleux O, Bastien F, Lamblin P, Pascanu R, Desjardins G, Turian J, Warde-Farley D, Bengio Y (2010) Theano: A cpu and gpu math expression compiler in python. In: Proceedings of Python for Scientific Computing Conference (SciPy), Austin, Texas, USA, pp 3–10

Burkett D, Petrov S, Blitzer J, Klein D (2010) Learning better monolingual models with unannotated bilingual text. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning, Uppsala, Sweden, pp 46–54

Cai S, Chiang D, Goldberg Y (2011) Language-independent parsing with empty elements. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, Portland, Oregon, pp 212–216

Chen C, Ng V (2013) Chinese zero pronoun resolution: Some recent advances. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, pp 1360–1365

Chung T, Gildea D (2010) Effects of empty categories on machine translation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, Massachusetts, USA, pp 636–645

Dagan I, Itai A, Schwall U (1991) Two languages are more informative than one. In: Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, Berkeley, California, USA, pp 130–137

Haspelmath M (2001) The European linguistic area: standard average European. In: Language typology and language universals. (Handbücher zur Sprach-und Kommunikationswissenschaft), vol 2, Berlin: de Gruyter, pp 1492–1510

Huang CTJ (1984) On the distribution and reference of empty pronouns. Linguistic Inquiry 15(4):531–574

Huang F, Vogel S (2002) Improved named entity translation and bilingual named entity extraction. In: Proceedings of Fourth IEEE International Conference on Multimodal Interfaces (ICMI), Pittsburgh, PA, USA, pp 253–258

Hwa R, Resnik P, Weinberg A, Cabezas C, Kolak O (2005) Bootstrapping parsers via syntactic projection across parallel texts. Natural Language Engineering 11(3):311–325

Koehn P, Schroeder J (2007) Experiments in domain adaptation for statistical machine translation. In: Proceedings of the 2nd Workshop on Statistical Machine Translation, Prague, Czech Republic, pp 224–227

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, pp 177–180

Kong F, Zhou G (2010) A tree kernel-based unified framework for chinese zero anaphora resolution. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, Massachusetts, USA, pp 882–891

Kuzman Ganchev JG, Taskar B (2009) Dependency grammar induction via bitext projection constraints. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, pp 369–377

Le Nagard R, Koehn P (2010) Aiding pronoun translation with co-reference resolution. In: Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden, pp 252–261

Li CN, Thompson SA (1989) Mandarin Chinese: A functional reference grammar. University of California Press, Oakland, California, USA

Lison P, Tiedemann J (2016) Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In: Proceedings of the 10th International Conference on Language Resources and Evaluation, Portorož, Slovenia

Liu Q, Tu Z, Lin S (2013) A novel graph-based compact representation of word alignment. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Sofia, Bulgaria, pp 358–363, URL `http://www.aclweb.org/anthology/P13-2064`

Liu Y, Xia T, Xiao X, Liu Q (2009) Weighted alignment matrices for statistical machine translation. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2, Singapore, pp 1017–1026

Mesnil G, He X, Deng L, Bengio Y (2013) Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France, pp 3771–3775

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, USA, pp 3111–3119

Moore R (2003) Learning translations of name-entity phrases from parallel corpora. In: Proceedings of Meeting of the European Chapter of the Association of Computational Linguistics (EACL), Budapest, Hungary, pp 253–258

Nakamura M (1987) Japanese as a pro language. The Linguistic Review 6:281–296

Novak M, Zabokrtsky Z (2014) Cross-lingual coreference resolution of pronouns. In: Proceedings of the 25th International Conference on Computational Linguistics, Dublin, Ireland, pp 14–24

Och FJ (2003) Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, Sapporo, Japan, pp 160–167

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Computational Linguistics 29(1):19–51

Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp 311–318

Pradhan S, Moschitti A, Xue N, Uryupina O, Zhang Y (2012) CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In: Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, Jeju Island, Korea, pp 1–27

Quirk R, Greebaum S, Leech G, Svartvik J (1985) A Comprehensive Grammar of the English Language, vol 9. New York: Longman

Raymond C, Riccardi G (2007) Generative and discriminative algorithms for spoken language understanding. In: Proceedings of 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, pp 1605–1608

Rosti AVI, Ayan NF, Xiang B, Matsoukas S, Schwartz RM, Dorr BJ (2007) Combining outputs from multiple machine translation systems. In: Proceedings of the Human Language Technology and the 6th Meeting of the North American Chapter of the Association for Computational Linguistics, Rochester, NY, USA, pp 228–235

Si L, Callan J (2005) Clef 2005: Multilingual retrieval by combining multiple multilingual ranked lists. In: Proceedings of Accessing Multilingual Information Repositories, Vienna, Austria, pp 121–130

Snyder B, Naseem T, Barzilay R (2009) Unsupervised multilingual grammar induction. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, Singapore, pp 73–81

Stolcke A (2002) Srilm - an extensible language modeling toolkit. In: Proceedings of the 7th International Conference on Spoken Language Processing, Colorado, USA, pp 901–904

Taira H, Sudoh K, 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, Jeju, Republic of Korea, pp 111–118

Tu Z, Liu Y, Hwang YS, Liu Q, Lin S (2010) Dependency forest for statistical machine translation. In: Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, pp 1092–1100

Tu Z, Liu Y, Liu Q, Lin S (2011) Extracting Hierarchical Rules from a Weighted Alignment Matrix. In: Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, pp 1294–1303

Tu Z, Liu Y, He Y, van Genabith J, Liu Q, Lin S (2012) Combining multiple alignments to improve machine translation. In: Proceedings of the 24rd International Conference on Computational Linguistics, Mumbai, India, pp 1249–1260

Wang L, Tu Z, Zhang X, Li H, Way A, Liu Q (2016a) 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, San Diego, California, USA, pp 983–993

Wang L, Zhang X, Tu Z, Li H, Liu Q (2016b) Dropped pronoun generation for dialogue machine translation. In: Proceedings of the IEEE International Conference of Acoustics, Speech and Signal Processing, Shanghai, China, pp 6110–6114

Wang L, Zhang X, Tu Z, Way A, Liu Q (2016c) Automatic construction of discourse corpus for dialogue translation. In: Proceedings of the 10th Language Resources and Evaluation Conference, Portorož, Slovenia, p 2748–2754

Wang M, Manning DC (2014) Cross-lingual projected expectation regularization for weakly supervised learning. Transactions of the Association of Computational Linguistics 2:55–66

Xiang B, Luo X, Zhou B (2013) Enlisting the ghost: Modeling empty categories for machine translation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, pp 822–831

Xu J, Deng Y, Gao Y, Ney H (2007) Domain dependent statistical machine translation. In: Proceedings of the MT Summit XI, Copenhagen, Denmark, pp 515–520

Xue N, Yang Y (2013) Dependency-based empty category detection via phrase structure trees. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, USA, pp 1051–1060

Xue N, Xia F, Chiou FD, Palmer M (2005) The Penn Chinese Treebank: Phrase structure annotation of a large corpus. Natural Language Engineering 11(02):207–238

Yang Y, Xue N (2010) Chasing the ghost: recovering empty categories in the Chinese treebank. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, Beijing, China, pp 1382–1390

Yang Y, Liu Y, Xu N (2015) Recovering dropped pronouns from Chinese text messages. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China, pp 309–313

Yarowsky D, Ngai G (2001) Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In: Proceedings of the 2nd meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL), Pittsburgh, PA, USA, pp 1–8

Zhang S, Ling W, Dyer C (2014) Dual subtitles as parallel corpora. In: Proceedings of the 10th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, pp 1869–1874

Zhao S, Ng HT (2007) Identification and resolution of Chinese zero pronouns: A machine learning approach. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, pp 541–550