Engaging Content
Engaging People

A Novel Approach to Dropped Pronoun Translation

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Hang Li, Qun Liu

Presented by Qun Liu

ADAPT Centre, Dublin City University

2016-05-18 Nanjing Normal University

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union

European Union
European Regional Development Fund

Science Foundation Ireland

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Outline

- **Motivation**
  - Dropped Pronoun in Machine Translation
  - Pronouns in English and Chinese
- **Related Work**
- **Methodology**
  - DP Training Corpus Annotation
  - DP Generation
  - Integrating into Translation
- **Experiments**
- **Conclusion**

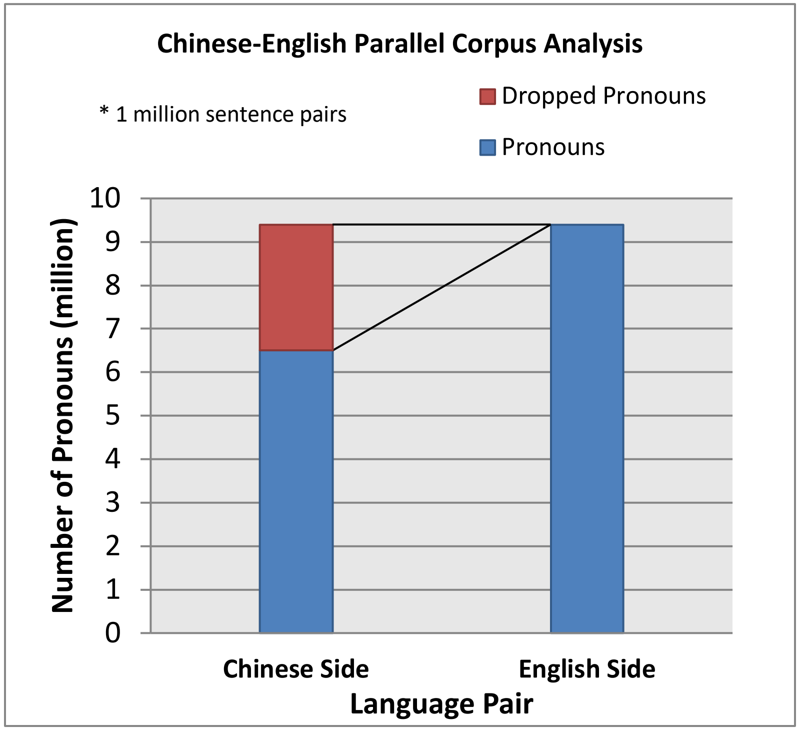
Dropped Pronoun in Machine Translation

Dropped pronouns (DPs) are challenges in **machine translation**, when certain classes of pronouns are frequently dropped in the **source language** but should retained in the **target language**.

- Pro-drop languages: **Chinese**, **Japanese**, Korean etc.
- Non-pro-drop languages: French, German, and **English** etc.

Dropped Pronoun in Machine Translation

This poses difficulties for Statistical Machine Translation (SMT) from **pro-drop languages** (e.g. Chinese) to **non-pro-drop languages** (e.g. English), since translation of such missing pronouns cannot be normally reproduced.



Pronouns in English and Chinese

Quirk et al (1985) classifies the principal English pronouns into three groups: **personal pronouns**, **possessive pronouns** and **reflexive pronouns**, called **central pronouns**.

In our work, we mainly focus on central pronouns in English-Chinese for MT.

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*)

* Correspondence of pronouns in Chinese-English (abbreviations: person type = 1st, 2nd, 3rd, singular = SG, plural = PL, male = M, female = F and neutral = N).

Related Work

There is some work related to DP generation:

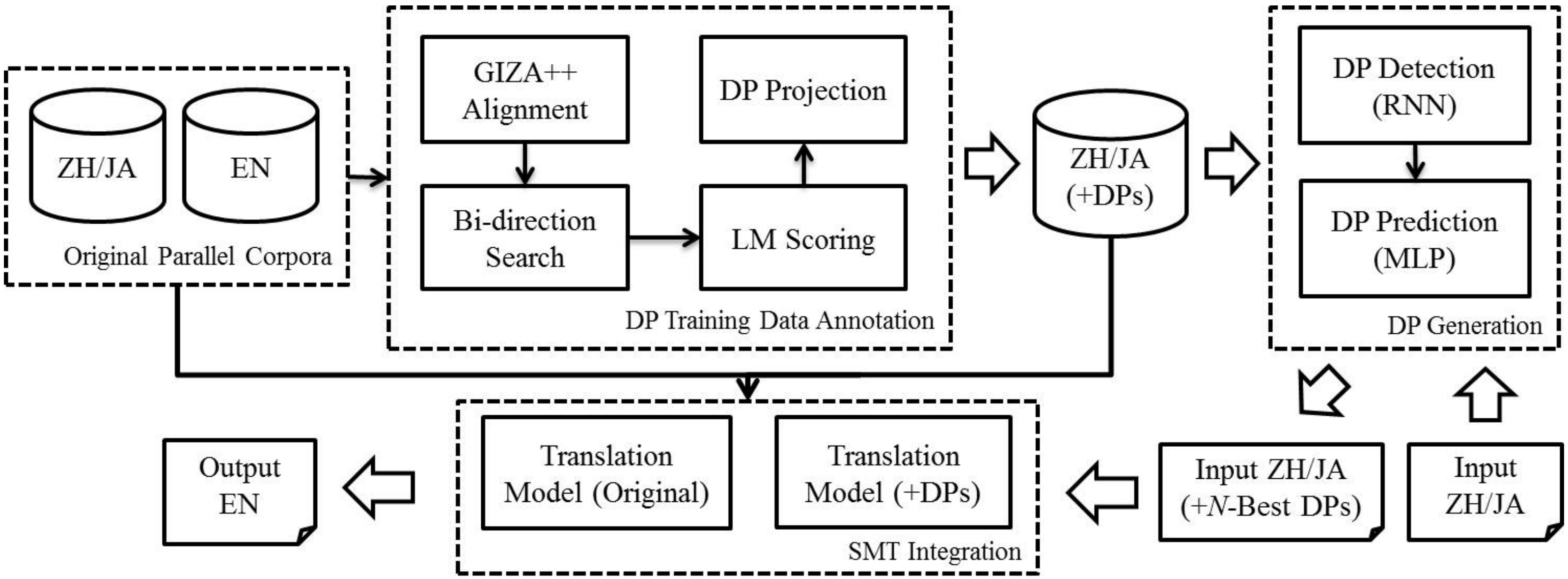
- Zero pronoun resolution, which is a sub-direction of co-reference resolution (Zhao and Ng, 2007; Kong and Zhou, 2010; Chen and Ng, 2013).
- Empty categories, which aims to recover long-distance dependencies, discontinuous constituents and certain dropped elements in phrase structure treebanks (Yang and Xue, 2010; Cai et al, 2011; Xue and Yang, 2013).

The above methods can also be used for DP translation using SMT:

- Taira et al (2012) propose both simple rule-based and manual methods to add zero pronouns on the source side for Japanese-English translation.
- Le Nagard and Koehn (2010) present a method to aid English pronoun translation into French for SMT by integrating co-reference resolution.

Methodology

We propose a universal **architecture** of our method, which can be divided into three main components: **DP training data annotation**, **DP generation**, and **SMT integration**.



DP Training Corpus Annotation

We propose **bidirectional search** method to automatically annotate DPs by utilizing alignment information.

We first algorithm to detect **possible positions** for DP.

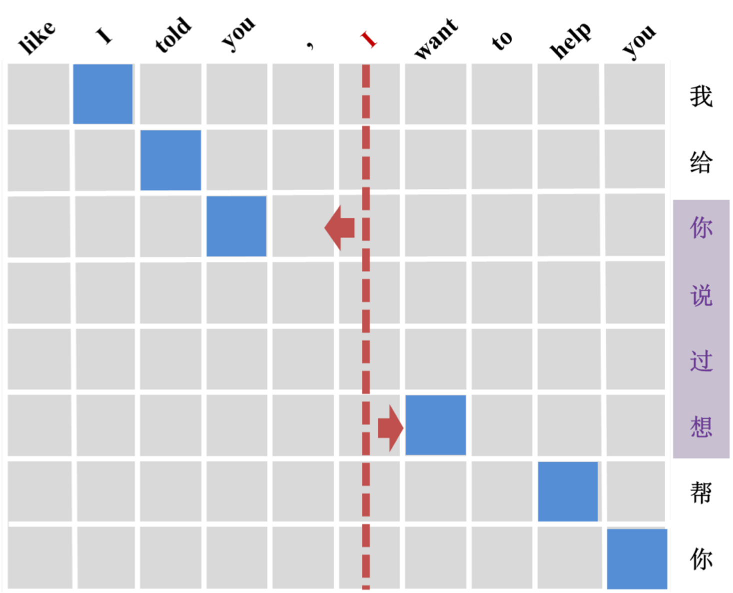
Algorithm 1 Bidirectional search algorithm in MATLAB™

```
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
```

To further determine the **exact position of DP**, we score all possible sentences with inserting corresponding Chinese DP using **language models** trained on a lager corpus.

DP Training Corpus Annotation

We use an example to illustrate our idea.

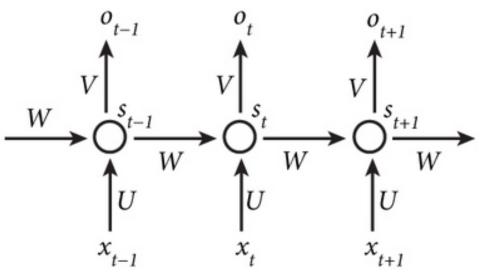


ID	possible positions to insert DP-I
1	我 给 你 **DP-I** 说 过 想 帮 你
2	我 给 你 说 **DP-I** 过 想 帮 你
3	我 给 你 说 过 **DP-I** 想 帮 你
4	我 给 你 说 过 想 帮 你

DP Generation

We parse this task into two phases: **DP detection** and **DP prediction**.

•**DP detection**. We employ RNN and regard it as sequence labelling problem.



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

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

$$y^{(t)} = g(W\mathbf{h}^{(t)})$$

•**DP prediction**. Based on detection results, we use a MLP with rich features: lexical, context and syntax.

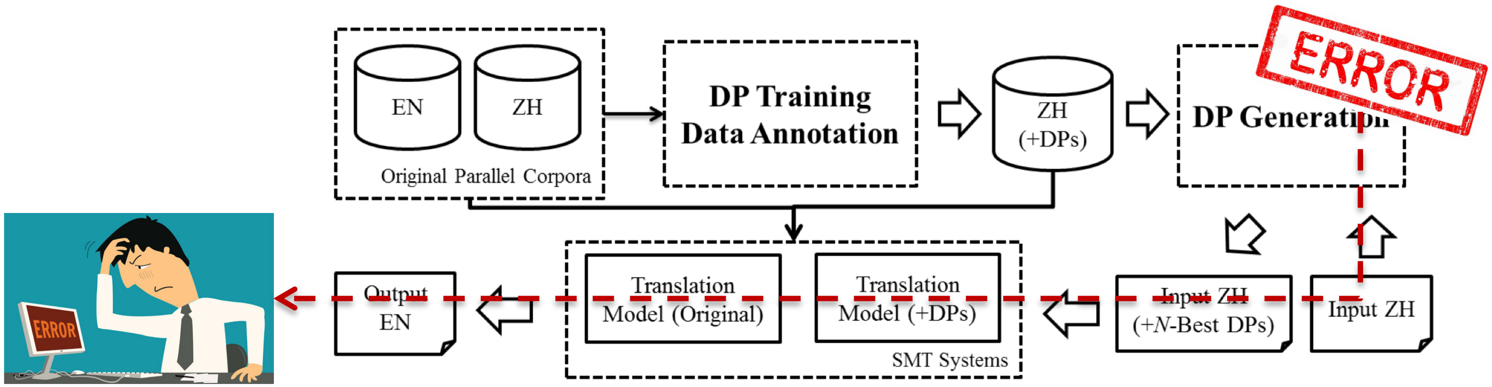
Actually, in our pilot experiments [3], we also employ **LMs** to select the best DP from all pronoun candidates. However, the performance is not good.

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

[3] Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Hang Li, Qun Liu. "Dropped Pronoun Generation For Dialogue Machine Translation." ICASSP. 2016.

Integrating into Translation

DP-inserted translation model (**DP-ins. TM**) and DP-generated input (**DP-gen. Input**).



But **DP-gen. Input** suffers from a major drawback: it only uses 1-best prediction result for decoding, which potentially introduces translation mistakes due to the propagation of prediction errors.

N-best DP-gen. Input. feed the decoder (via confusion network decoding) N-best prediction results, which allows the MT to arbitrate between multiple ambiguous hypotheses.

Experiments

For training data, we extract around **1M sentence pairs** (movie or TV episode subtitles) from subtitle websites.

- keep contextual information.
- manually create development and test sets.
- two LMs for the **DP annotation** and **translation tasks**, respectively.

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

- phrase-based SMT model in Moses; 5-gram language models using the SRI Language Toolkit; GIZA++; minimum error rate.
- case-insensitive NIST BLEU.
- Theano neural network toolkit to implement RNN and MLP.

Results - DP Annotation

- We first check whether the DP annotating strategy is reasonable.
- We **automatically** and **manually** insert DPs into the source sides of development and test data with considering their target sides.
- The agreements between automatic labels and manual labels are:
  - ❑ DP detection: **94%** and **95%** on development set and test set;
  - ❑ DP prediction: **92%** and **92%** on development set and test set.
- This indicates that the automatic annotate strategy is **trustworthy** for DP generation and DP-inserted translation model.

Results - DP Generation

We then measure the accuracies (in terms of words) of our generation models in two phases: **DP detection** and **DP prediction**.

- **DP Detection ("Position")**. We only consider the tag for each word (drop or not drop before the current word), without considering the exact pronoun for DPs.
- **DP Prediction ("+Pronouns")**. We consider both the DP position and predicted pronoun.

DP	Set	P	R	F1
DP Detection	Dev	0.88	0.84	0.86
	Test	0.88	0.87	0.88
DP Prediction	Dev	0.67	0.63	0.65
	Test	0.67	0.65	0.66

Table 3: Evaluation of DP generation quality.

Results - MT Integration

- **Baseline** are relatively low because 1) only one reference and 2) dialogue domain.
- **+DP-ins. TM** indicates that the DP insertion is helpful to alignment.
- **+DP-gen. Input N** is a more soft way of integration than 1-best.
- **Oracle** shows that there is still a large space of improvement for the DP generation model.

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)

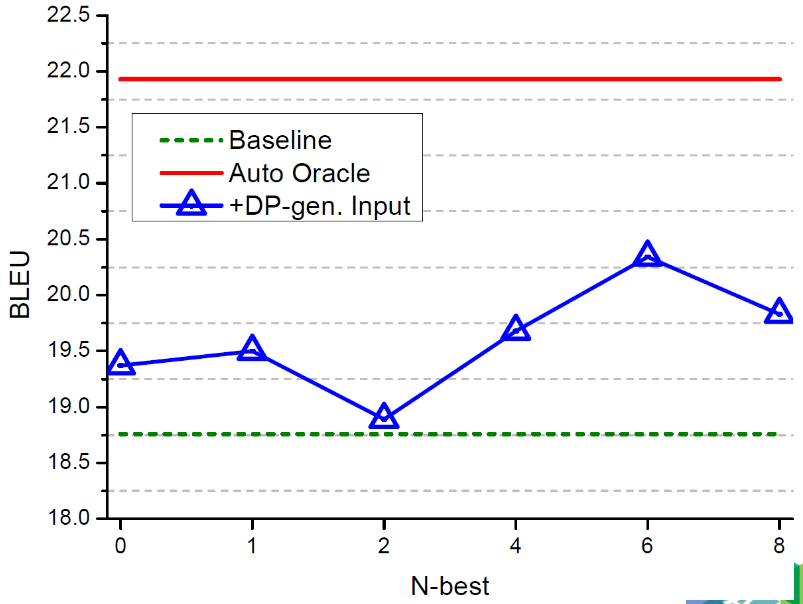


Table 4: Evaluation of DP translation quality.

Analysis

Case A (Better)

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

(1-best)
〈你〉 想不想 听 一件 奇怪的 事 ？
Do 〈**you**〉 want to hear something weird ?

(reference) Do you want to hear something weird ?

Case B (Unchanged)

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

(1-best)
不要 告诉 瑞秋 ， 〈你〉 待会 见 。
Do not tell Rachel . See 〈**you**〉 later .

(reference) Do not tell Rachel . See you later .

Case C (Worse)

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

(1-best)
你 肯定 〈我〉 看过 那 电视剧。
You are sure 〈**I**〉 ’ve seen that show .

(reference) You must have seen that one .

Case D

(Baseline)
都 不会 想 我 吗 ？
Won ‘t even miss me ?

(1-best)
〈我〉 都 不会 想 我 吗 ？
〈**I**〉 won ‘t even miss me ?

(2,4,6-best)
〈我|你|...〉 都 不会 想 我 吗？
You won ‘t even miss me ?

(8-best)
〈我|你|他|...〉 都 不会 想 我 吗？
He won ‘t even miss me ?

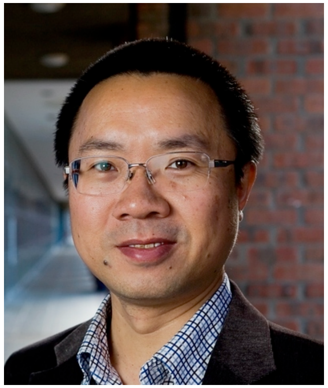
(reference) You won ‘t even miss me ?

Conclusion and Future Wok

- Bilingual information is helpful to set up a monolingual model without any manually annotated training data;
- Benefited from representation learning, NN-based models work well without complex feature engineering work;
- N-best (a soft way) DP integration works better than ponderous 1-best insertion.

In future work, we plan to extend our work to different genres, languages and other kinds of dropped words to validate the robustness of our approach.

Q&A



Qun Liu, Professor, Dr.,PI

ADAPT Centre
Dublin City University

Institute of Computing Technology
Chinese Academy of Sciences

Email: qun.liu@dcu.ie