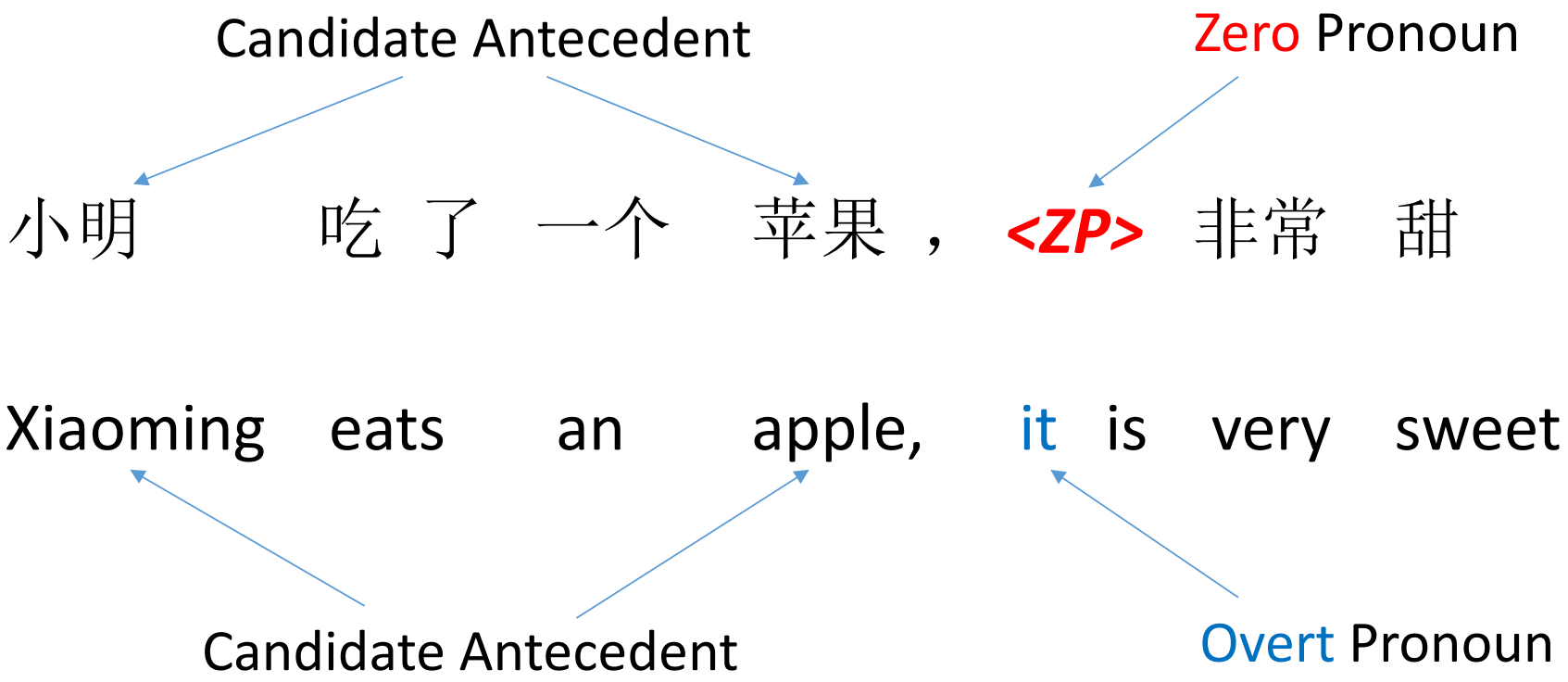
Generating and Exploiting Large-scale Pseudo Training Data for Zero Pronoun Resolution

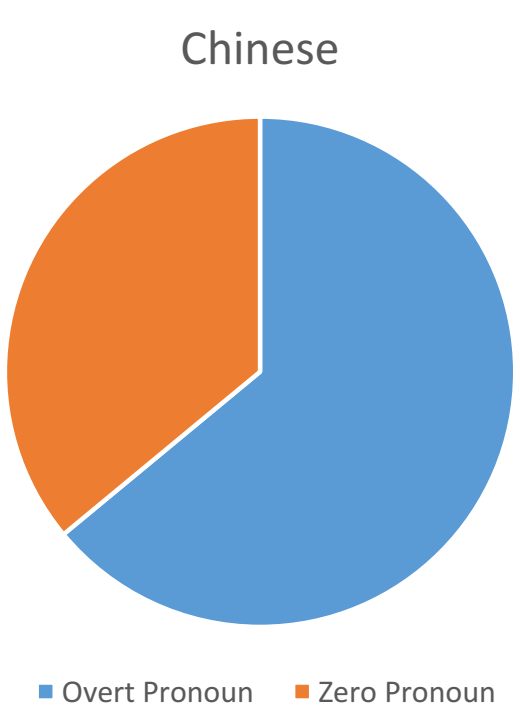
Ting Liu[†], Yiming Cui[‡], Qingyu Yin[†], **Weinan Zhang**[†],
Shijin Wang[‡] and Guoping Hu[‡]

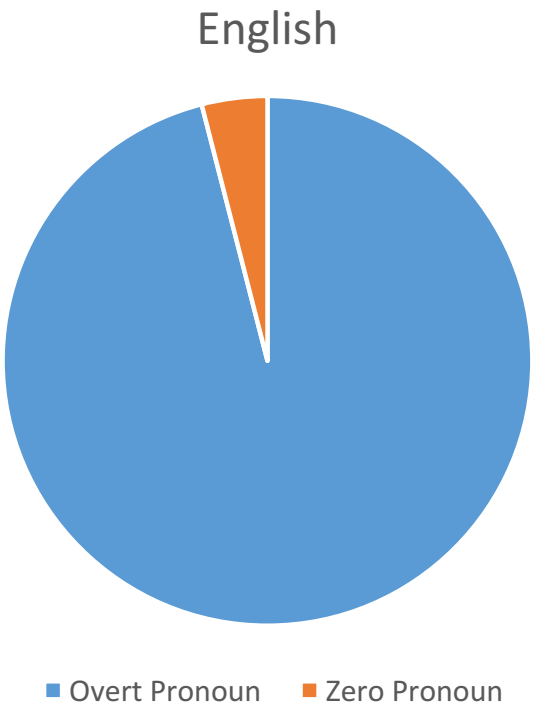
[†]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
[‡]iFLYTEK Research, Beijing, China

Zero Pronoun (ZP)



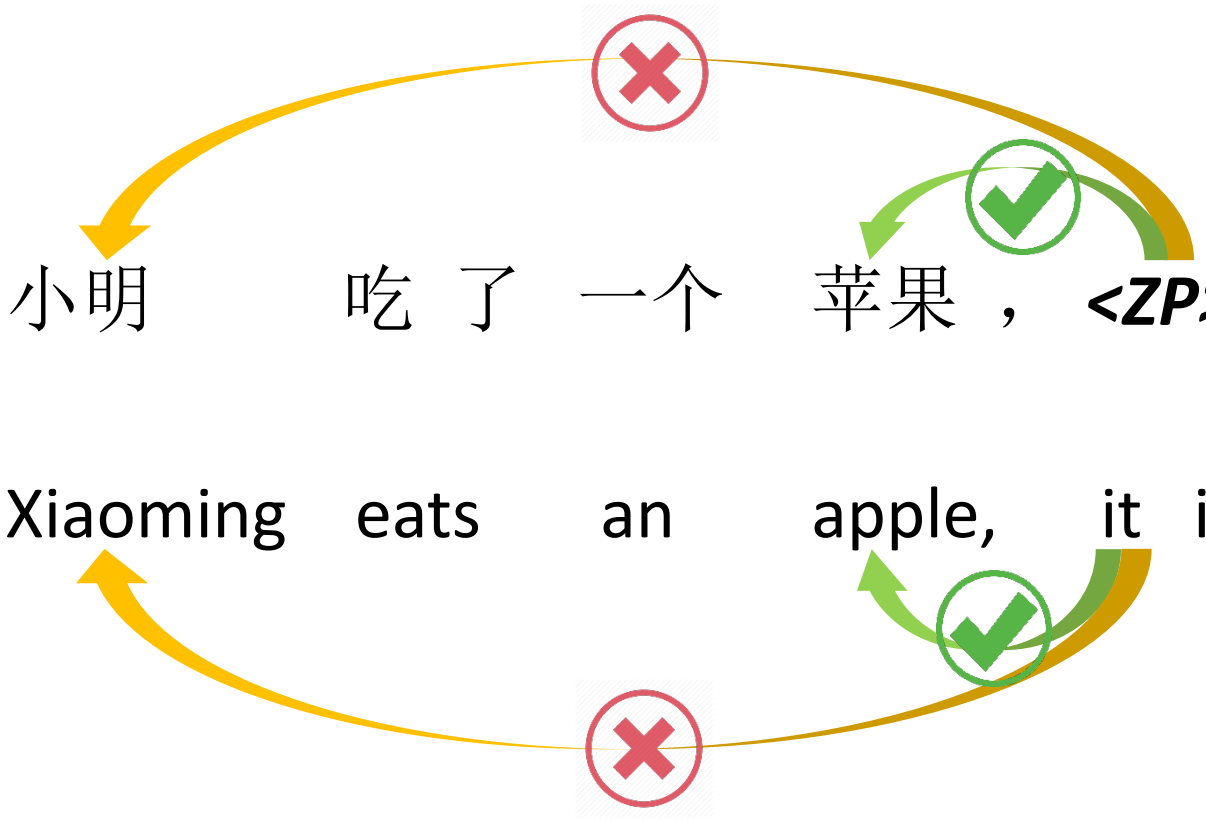
ZP Proportion





[1]Kim Y J. Subject/object drop in the acquisition of Korean: A cross-linguistic comparison[J]. Journal of East Asian Linguistics, 2000, 9(4): 325-351.
[2]Zhao S, Ng H T. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach[C]// EMNLP-CoNLL 2007.

Zero Pronoun Resolution (ZPR)



Challenges of ZPR

- No overt pronoun for indication
  - No information for the positions of ZPs
  - No type/surface information of ZPs

- Feature engineering

Syntactic features (13)	whether z is the first gap in an IP clause; whether z is the first gap in a subject-less IP clause, and if so, POS(w_1); whether POS(w_1) is NT; whether w_1 is a verb that appears in a NP or VP; whether P_l is a NP node; whether P_r is a VP node; the phrasal label of the parent of the node containing POS(w_1); whether V has a NP, VP or CP ancestor; whether C is a VP node; whether there is a VP node whose parent is an IP node in the path from w_1 to C.
Other features (6)	whether z is the first gap in a sentence; whether z is in the headline of the text; the type of the clause in which z appears; the grammatical role of z (SUBJECT, OBJECT, or OTHER); whether w_{-1} is a punctuation; whether w_{-1} is a comma.

19 hand-crafted features for ZP

Syntactic features (12)	whether c has an ancestor NP, and if so, whether this NP is a descendent of c's lowest ancestor IP; whether c has an ancestor VP, and if so, whether this VP is a descendent of c's lowest ancestor IP; whether c has an ancestor CP; the grammatical role of c (SUBJECT, OBJECT, or OTHER); the clause type in which c appears; whether c is an adverbial NP, a temporal NP, a pronoun or a named entity.
Distance features (4)	the sentence distance between c and z; the segment distance between c and z, where segments are separated by punctuations; whether c is the closest NP to z; whether c and z are siblings in the associated parse tree.
Other features (2)	whether c is in the headline of the text; whether c is a subject whose governing verb is lexically identical to the verb governing of z.

18 hand-crafted features for antecedent

Chen Chen and Vincent Ng. 2016. Chinese zero pronoun resolution with deep neural networks. ACL 2016.

Solutions

- No overt pronoun for indication
  - Considering all possible positions for ZPs identification
  - Classifying ZPs to Anaphoric ZPs (AZP) and Non-AZPs
  - Modelling the semantics of ZPs and antecedents

Most existing work

- Feature engineering
  - Automatically learning to represent features
  - Deep learning approaches for the modeling
  - More labeled data for training

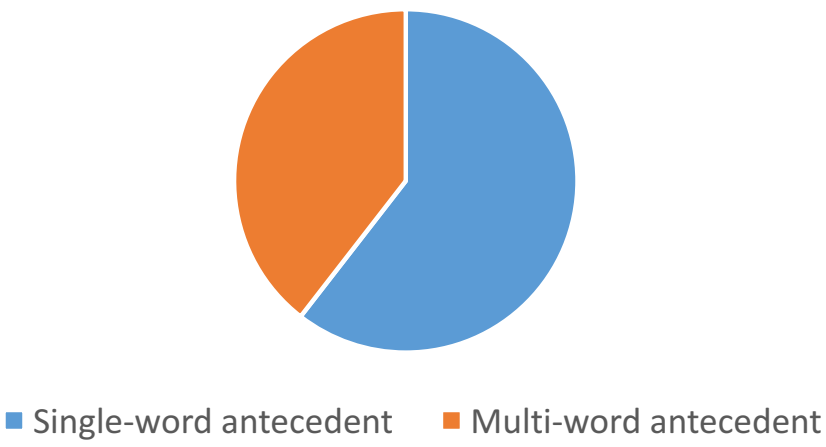
This paper

How to Obtain Large-scale Training Data?

- Manual Annotation
  - Labor consuming
  - Hard to say “large-scale”

- Automatic Generation
  - Easy to obtain
  - Large-scale
  - Pseudo training data

What is Actual Training Data?



- Sample Training Data in OntoNotes 5.0
  - Single-word (In Chinese) antecedent

CN：[警方] 怀疑 这是 一起 黑枪 案件，**zp$_1$** 将 枪械 交送 市里 **zp$_2$** 以 清理 案情 。
EN：[*The police*] *suspected that this is a criminal case about illegal guns,* **zp$_1$** *brought the guns to the city* **zp$_2$** *to deal with the case.*

  - Multi-word antecedent

CN：这次 [近 50 年 来 印度 发生 的 最 强烈 地震] 震级 强，**zp** 波及 范围 广，印度 邻国 如 尼泊尔 也 受到 了 影响 。
EN：[*The earthquake that is the strongest one occurs in India within recent 50 years*] *has a high-magnitude,* **zp** *influences a large range of areas, and the neighboring country of India like Nepal is also affected.*

How to Generate Pseudo Training Data?

- Collecting large-scale news documents, which is relevant (or homogenous in some sense) to the OntoNotes 5.0 data.
- Given a document $\mathcal{D}$, a word is randomly selected as an answer $\mathcal{A}$ if
  - It is either a noun or pronoun
  - It should appear at least twice in the document
- The sentence contains $\mathcal{A}$ is defined as a query $\mathcal{Q}$, in which the answer $\mathcal{A}$ is replaced by a specific symbol "**<blank>**"

Document:

1 ||| welcome both of you to the studio to participate in our program ,
欢迎 两 位 呢 来 演播室 参与 我们 的 节目 ,

2 ||| it happened that i was going to have lunch with a friend at noon .
正好 因为 我 也 和 朋友 这个 , 这个 中午 一起 吃饭 。

3 ||| after that , i received an sms from 1860 .
然后 我 就 收到 1860 的 短信 。

4 ||| uh-huh , it was by sms .
嗯 , 是 通过 短信 的 方式 ,

5 ||| uh-huh , that means , er , you knew about the accident through the source of radio station .
嗯 , 就是说 呃 你 是 通过 台 里面 的 一 个 信息 的 渠道 知道 这儿 出 了 这样 的 事故 。

6 ||| although we live in the west instead of the east part , and it did not affect us that much ,
虽然 我们 生活 在 西部 不 是 在 东部 , 对 我们 影响 不 是 很 大 ,

7 ||| but i think it is very useful to inform people using sms .
但是 呢 , 我 觉得 有 这样 一 个 短信 告诉 大家 呢 是 非常 有用 的 啊 。

Query:

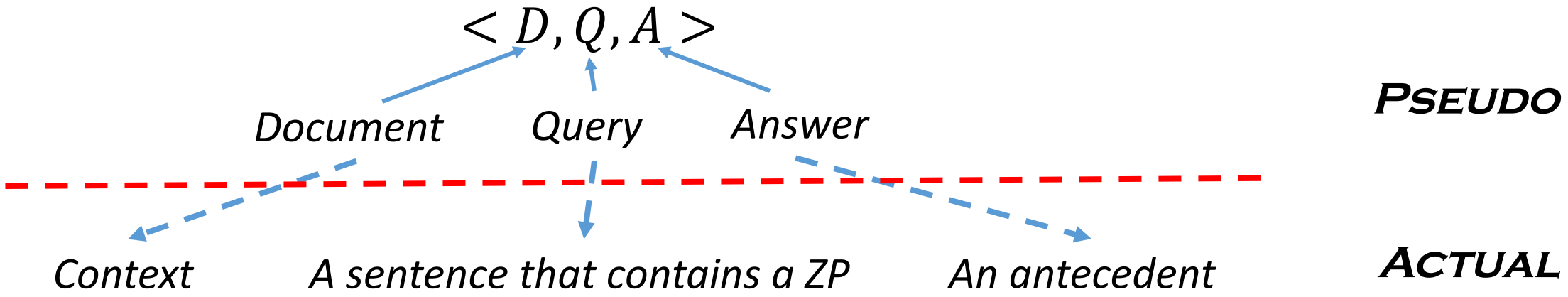
8 ||| some car owners said that **<blank>** was very good。
有 车主 表示 , 说 这 **<blank>** 非常 的 好。

Answer:

sms
短信

Zero Pronoun Resolution (ZPR)

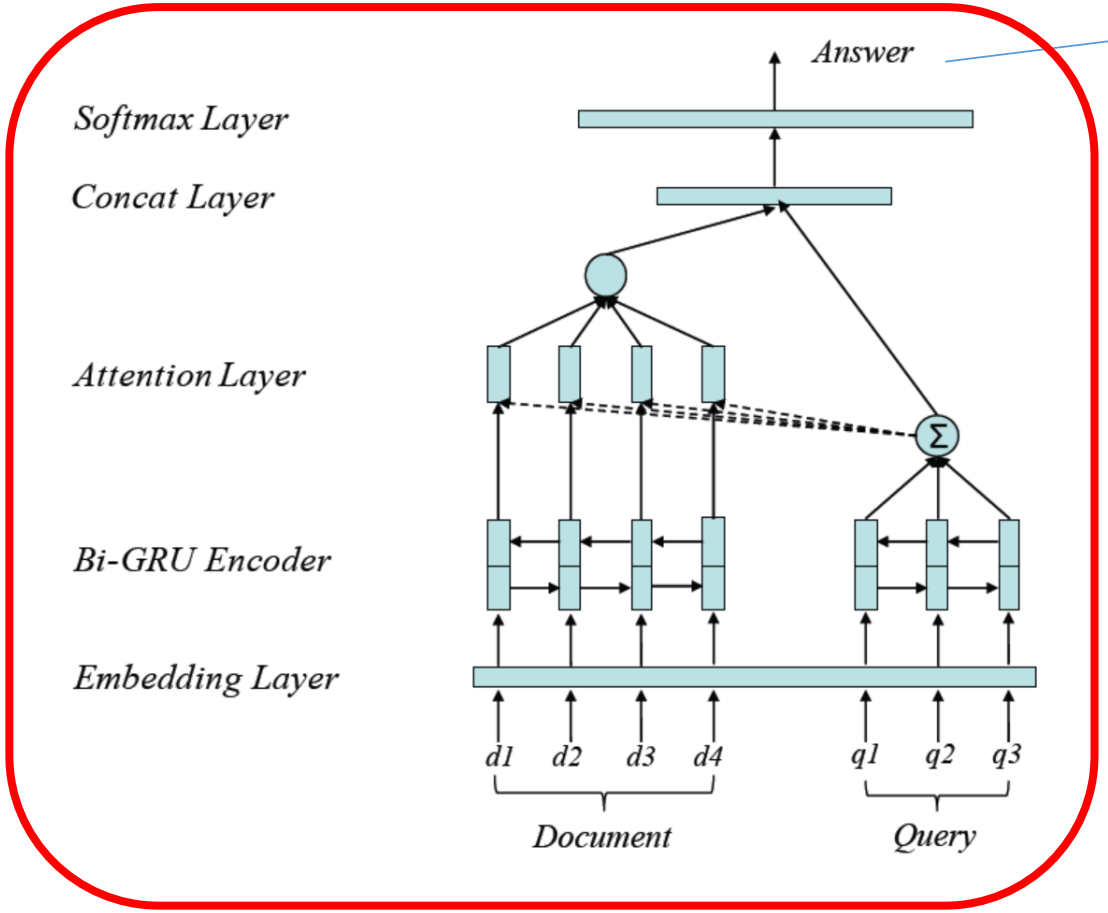
- A pseudo training sample can be represented as



- Zero pronoun resolution task is thus defined as

$$P(A|D,Q)$$

Attention-based NN Model for ZPR



Single-word

Matching the single word

Matching the head word

Single-word Antecedent

Multi-word Antecedent

Two-step Training

Pseudo Data Pre-training

Actual Data Fine-tuning

OR

General Training

Domain Training

Experimental Data

- OntoNotes Release 5.0 from CoNLL-2012
  - Broadcast News (BN), Newswires (NW), Broadcast Conversations (BC), Telephone Conversations (TC), Web Blogs (WB), Magazines (MZ)

PSEUDO

	Sentences #	Query #
General Train	18.47M	1.81M
Domain Train	122.8K	9.4K
Validation	11,191	2,667

ACTUAL

	Docs	Sentences	Words	AZPs
Test	172	6,083	110K	1,713

Overall Performance

- F-score

	NW (84)	MZ (162)	WB (284)	BN (390)	BC (510)	TC (283)	**Overall**
Kong and Zhou (2010)	34.5	32.7	45.4	51.0	43.5	48.4	44.9
Chen and Ng (2014)	38.1	31.0	50.4	45.9	53.8	**54.9**	48.7
Chen and Ng (2015)	46.4	39.0	51.8	53.8	49.4	52.7	50.2
Chen and Ng (2016)	48.8	41.5	56.3	**55.4**	50.8	53.1	52.2
Our Approach†	**59.2**	**51.3**	**60.5**	53.9	**55.5**	52.9	**55.3**

Effect of UNK Processing

(a) *The weather today is not as pleasant as the weather of yesterday.*

(b) *The* ***<unk>*** *today is not as* ***<unk>*** *as the* ***<unk>*** *of yesterday.*

(c) *The* ***<unk1>*** *today is not as* ***<unk2>*** *as the* ***<unk1>*** *of yesterday.*

	F-score
Without UNK replacement	52.2
With UNK replacement	**55.3**

Effect of Domain Adaptation

	F-score
Only Pseudo Training Data	41.1
Only Task-Specific Data	44.2
Only Task-Specific Data + GloVe	50.9
Domain Adaptation	**55.3**

Error Analysis

- The impact of UNK words

CN： **zp** unk1 unk2 顶，将 unk3 和 unk4 的 美景 尽收眼底 。
EN： **zp** *successfully [climbed]$_{unk1}$ the peak of [Taiping Mountain]$_{unk2}$, to have a panoramic view of the beauty of [Hong Kong Island]$_{unk3}$ and [Victoria Harbour]$_{unk4}$.*

- Long distance between ZPs and antecedents

CN： [我] 帮 不 了 那个 人… （多于30个词）… 那 天 结束 后， **zp** 回到 家 中 。
EN： *[I] can't help that guy … (more than 30 words) … After that day,* **zp** *return home.*

Conclusion

- Generating and exploiting pseudo training data for ZPR
  - Inspired by the cloze-style reading comprehension
- Two-step training of the ZPR model for the use of the large scale pseudo training data
- A new State-of-the-Art approach on Chinese ZPR task

Thanks!
Questions and Advices?