Zero Pronoun Identification in Chinese Language with Deep Neural Networks

Tao Chang, Shaohe Lv, Xiaodong Wang and Dong Wang
National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, Changsha, China

***Abstract*—Zero pronoun resolution is very important in natural language processing. Identification of zero pronoun is the fundamental task of its resolution. Existing feature engineering based identification approaches are unsuitable for big data applications due to labor-intensive work. Furthermore, as extracted from parse trees which are not unique for a certain sentence, features may be improper for zero pronoun identification. In this paper, we constructed a two-layer stacked bidirectional LSTM model to tackle identification of zero pronoun. To extract semantic knowledge, the first layer obtains the structure information of the sentence, and the second layer combines the part-of-speech information with obtained structure information. The results in two different kinds of experimental environment show that, our approach significantly outperforms the state-of-the-art method with an absolute improvement of 4.3% and 20.3% F-score in OntoNotes 5.0 corpus respectively.**

Keywords-zero pronoun; Identification; LSTM

I. Introduction

Coreference resolution is very important in natural language processing (NLP), which is a task of determining whether two or more noun pronoun phrases refer to the same entity in a text. Proper coreference resolution will establish the connection between the anaphor entity and the following part behind the noun phrase, which is very useful to build effective language modeling. Moreover, it will benefit other NLP applications such as discourse analysis, information extraction etc.

Zero pronoun (ZP) resolution is a special kind of coreference resolution. A ZP is a gap in a sentence that should have a noun pronoun. ZP resolution is the task of finding where the gap is and determining which entity the gap refers to. Due to the gaps, the language model may obtain some error knowledge of the context adjacent to the gap. Thus, ZP resolution can not only achieve the same effect with the coreference resolution, but also be benefit to properly model the contextual information.

The ZP resolution can be implemented through two steps, to locate the gap firstly, and then to determine which entity it refers to. If the position of a ZP is determined, the ZP resolution can be reduced to a common coreference resolution. Therefore, identification of ZP is crucial. However, this task has not been paid enough attention to, as the recent work [1][2][3][4] just focused on resolving ZPs which have been annotated manually. To our best knowledge, only [5][6][7] made some progresses in identification.

Those models aforementioned have several weaknesses. Firstly, most of those previous identification approaches were based on feature engineering, which were inefficient and incapable to handle big data. Secondly, most of those models were only implemented with syntactic and lexical features, which deeply relied on parse trees. For example, the F-score of Chen et al.'s system [2] dropped from 60.1% to 36.1% when using automatic parse trees. Moreover, we argue that semantic information is the internal connection between the ZP and its context. However, it is hard to model semantic features within feature engineering based approaches.

In this paper, a two-layer stacked bidirectional Long Short-Term Memory (LSTM) model is proposed to tackle ZP identification. To extract semantic knowledge, the first layer obtains the structure information of the sentence, and the second layer combines the part-of-speech information with obtained structure information. As ZP occurs much more frequently in Chinese compared to English [8], this paper focuses on identification in Chinese language.

Firstly, deep architecture learning can automatically discover deep abstractions, from lower level features to higher level concepts [9]. Through the deep networks, the features which make use of detecting ZP can be automatically obtained and kept in the hidden units. Thus, the identification can start from raw text without features extracting manually. Secondly, the LSTM can recursively compose each word in a sentence, and generate a sequence of meaningful representations till the current word. This kind of order-sensitive model can capture the semantics of natural language, which is the key point of the identification. Thirdly, as the part-of-speech (POS) indicates the usage of a word, we extensively integrate the POS feature to obtain better semantic information representation.

In this paper, we build a model that can not only capture ZP contextual to get the semantic representations, but also integrate with the POS tags as an auxiliary feature. When evaluated on the Chinese portion of OntoNotes 5.0 with two different kinds of experimental settings, our approach outperforms the state-of-the-art method significantly with an absolute improvement of 4.3% and 20.3% F-score respectively.

The rest of this paper is organized as follows. Section 2 briefly describes the related work on ZP identification. Section 3 introduces the Chinese ZP identification. Section 4 presents the details of our approach. Section 5 reports the experimental results and analyzes the errors. Finally, we conclude our work in section 6.

II. Related Work

This section briefly overviews the related work on ZP identification.

In [5], ZP resolution was firstly divided into two subtasks: identification and resolution. A heuristic rule was employed to detect the ZP. This rule could recover all zero anaphors, but suffered from low precision by introducing too many false zero anaphors and therefore leaded to relatively poor performance in anaphoricity determination.

In [6], a tree kernel-based ZP detection was proposed. For a given zero anaphor candidate, a proper parse tree structure must be constructed by first keeping the path from the root node to the predicate phrase node and then attaching all the immediate verbal phase nodes and nominal phrase nodes.

In [7], a unified framework was proposed to recover empty categories for Chinese given skeletal parse tree as input, and the empty categories were detected via manual features. Although the performance of their model sounded well, but only 931 files included in their corpus, moreover these files were specially selected.

In [10] [2], compound features proposed both by Zhao et al. and Yang et al. were employed to detect the ZP in OntoNotes 4.0 and OntoNotes 5.0. The identification performance was proved being strongly depended on the parse tree.

III. Problem Definition

A ZP is a gap in a sentence, which refers to an entity that supplies the necessary information for interpreting the gap.

Below is a typical example taken from the OntoNotes 5.0, which is translated from Chinese aligned. In the example, the ZPs (denoted as *pro*) refer to AIDS and Africa, respectively.

It is understood that the current spread of AIDS in Africa is very fast, the scope of the impact of *pro* is very large, if *pro* does not take effective preventive measures, Africa is the next 20 years, the most serious areas of AIDS raging.

From this example, we can see that a ZP only occurs between two words or before the first word, and does not occur after the last punctuation or word.

Inspired by the dropped pronoun detection approach of Wang et al. [11], the ZP detection is to label words if there are pronouns missing before a word, which can intuitively be modeled as a sequence labeling task.

Given an input sentence consisting of *n* words, we expect the output to be a sequence of labels. Each label indicates if a ZP exists before the corresponding word. In our task, there are two labels {ZP, NZP}, corresponding to ZP or none of ZP. We can use a code scheme for the labels, where y = 1 for the ZP label and y = 0 for NZP label. Thus, ZP identification can be modeled as a sequence binary classification.

ZP identification is relatively difficult due to following reasons: (1) Only about 4% words preceded by ZPs, which leads to sharply unbalance between positive and negative samples, (2) in Chinese both the subject and object can be omitted, which leads to multiple kinds of ZP, (3) a ZP may occur in any position of a sentence, which leads to higher complexity in its identification.

IV. Proposed Approach

A. Overview of Our Approach

In our work, we argue that both the forward and backward information are useful for detecting ZPs. Moreover, according to theoretical conclusion from [9], a deep hierarchical model can be more efficient in representing some functions than a shallow one. Empirical performance improvement is also observed in Long Short-Term Memory (LSTM) network compared with the shallow one [12].

As mentioned above, we develop a two-layer stacked bidirectional LSTM model for ZP identification. An illustration of our model is given in Figure I, more details will be presented in the next subsection. Input of the model consists of two related parts: one is the one-hot representation of words and another is the corresponding POS tags. The bidirectional LSTM (bLSTM) of the first layer captures some shallow features from the raw text. The bLSTM of the second layer extracts the higher-level abstraction with the output of the first layer and the POS tags. Then ZPs can be predicted through the output of the second layer.

Since words will be processed recurrently in LSTM, outputs corresponding to each word can be used to predict the corresponded ZPs. As another option, all ZPs can be predicted via outputs of all words. Experimental results show that prediction from all outputs performs better. We think the LSTM is capable to handle the whole contextual information, but may *forget* some long distance dependent information. Thus, the final identification result is obtained through a Multilayer Perceptron (MLP) with all outputs of stacked bLSTM.

B. Network Architecture

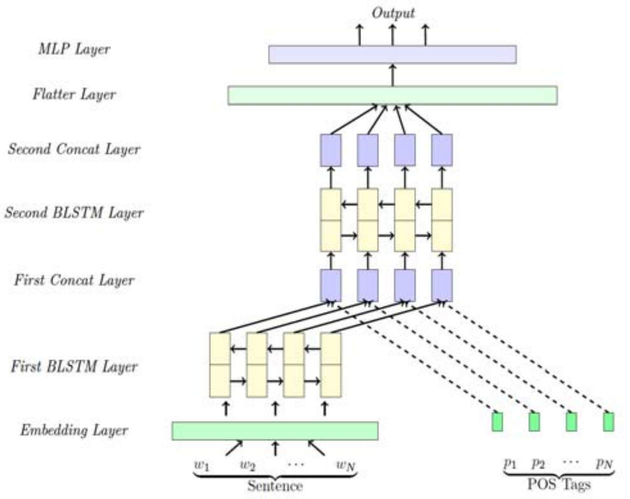

FIGURE I. ARCHITECTURE OF STACKED BIDIRECTIONAL LSTM MODEL FOR ZERO PRONOUN IDENTIFICATION

Our model is primarily a stacked bidirectional LSTM, which is showed in Figure I.

Firstly, the one-hot representation of each word in a sentence is projected into a low dimensional, continuous and real-valued vector with a shared embedding matrix $L^{d\times|V|}$, where *d* is the dimension of word vector and |*V*| is the size of word vocabulary. Then the embeddings are input into the bLSTM of the first layer to obtain their structure information.

$$e(x_i) = L_w \times w^i, where\ i \in [1, n] \tag{1}$$

$$\overrightarrow{h_i} = LSTM\big(e(x_1), \cdots, e(x_i)\big) \tag{2}$$

$$\overleftarrow{h_i} = LSTM\big(e(x_n), \cdots, e(x_i)\big) \tag{3}$$

where the LSTM network are implemented following [13] [14] [15]:

$$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t\text{-}1} + b_i) \\ \tilde{C}_t &= \tanh\,(W_c x_t + U_f + b_c) \\ f_t &= \sigma(W_f + U_f h_{t\text{-}1} + b_f) \\ C_t &= i_t{*}\tilde{C}_t + f_t{*}C_{t\text{-}1} \\ o_t &= \sigma(W_o x_t + U_o h_{t\text{-}1} + V_o C_t + b_o) \\ h_t &= o_t{*}\tanh\,(C_t) \end{aligned}$$

Secondly, in order to preserve more information for processing in next step, the output of both directional LSTM and the one-hot representation of POS tag vector are simply stacked together.

$$h_i = concat(\overrightarrow{h_i}, \overleftarrow{h_i}, p_i) \tag{4}$$

Thirdly, the merged vectors are input into the second layer bLSTM to capture higher level abstraction. Then the bidirectional outputs are stacked.

$$\overrightarrow{h_i'} = LSTM(h_1, \cdots, h_i) \tag{5}$$

$$\overleftarrow{h_i'} = LSTM(h_n, \cdots, h_i) \tag{6}$$

$$h_i' = concat(\overrightarrow{h_i'}, \overleftarrow{h_i'}) \tag{7}$$

At last, the identification result is predicted via a MLP with *softmax* activation function. As the MLP input and output are vectors, we fix the input length.

$$h_{flat} = flatten(h_1', \cdots, h_n') \tag{8}$$

$$y_i = softmax(W_i^o \times h_{flat} + b_i^o) \tag{9}$$

where y_i is a float number between 0 and 1. The final results are classified into two classes according to a pre-defined threshold *t*.

V. Experimental Result

In this section, we detail the experimental settings and the model implementation, analyze the results finally.

A. Experimental Settings

In this work, the Chinese portion of the OntoNotes 5.0 corpus is employed, which contains about 2,000 documents and over 1,000,000 tokens. The corpus includes six sources, namely Broadcast News (BN), Newswire (NW), Broadcast Conversation (BC), Telephone Conversation (TC), Web Blog (WB) and Magazine (MZ). In the corpus, a ZP is marked as *pro* or *PRO*. Statistics on the datasets are shown in Table I.

TABLE I. STATISTICS ON THE TRAINING AND TEST SETS

	Document	*Sentences*	*Token*	*Words*	*ZPs*
Training	1825	46102	1019757	49770	41420
Test	177	5129	113703	13279	4494

Chen et al.'s [2] system is employed as the baseline system, which gets the state-of-the-art results in Chinese ZP identification on Ontonotes 5.0 corpus. To evaluate our proposed approach, several experiments are conducted under two experimental settings. The experimental settings are the same with Chen et al. In Setting 1, golden syntactic parse trees are available. In Setting 2, system syntactic parse trees, which are obtained by Berkeley parser, are employed. The Berkeley parse is the state-of-the-art parsing model. The POS tags are obtained from corresponding parse trees.

The results of ZP identification are expressed in terms of recall (R), precision (P) and F-score (F), which is the same with Chen et al. [2].

B. Implementation Details

Our model is trained in a supervised end-to-end learning framework, and implemented with Theano [16] and Keras [17]. Training details are as follows.

- Training Data: Each training instance corresponds to a text sequence, a POS tag sequence and a ZP tag sequence.

  ■ The text sequence is the one-hot representation of a natural sentence with *all punctuations*.

  ■ The POS tag sequence is the one-hot representation of POS tags corresponding to the text sequence.

  ■ The ZP tag sequence is a "0-1" vector that each element represents the occurrence of a ZP. As mentioned, we have defined 1 for a ZP before the corresponding word, 0 for none.

- Loss Function: In our work, binary-cross-entropy error between standard classification and predicted classification is regarded as the loss function.
- Optimization: Root mean square propagation [18] (RMSprop) update rule is employed with initial learning rate of 0.01.
- Embedding: 200-dimension pre-trained word embedding is applied, which is obtained by training *word2vec* [19] on the Chinese portion of OntoNotes 5.0 corpus and Chinese Wikipedia [20].

- Output Threshold: As each output y_i of the model is a float number, which can be classified into 0 or 1 with a threshold t. In order to set the t, the model is trained in two phases. Firstly, 5% of the training data is preserved as the validation set, different t is tested on the validation part. As a result, F-score reaches the peak when t=0.35. And then, the model is trained with all training data and employed t=0.35 as the threshold.

C. Results Analysis

Experimental results of our approach and the baseline system are compared in Table II. As we can see, our approach significantly outperforms the baseline system under two experimental settings by 4.3% and 20.3% in terms of overall F-score, respectively. Especially, our model completely outperforms the baseline under Setting 2, due to the POS tag is used only as an assistant feature.

TABLE II. EXPERIMENTAL RESULTS ON TEST DATA

	Setting1: Gold Parse			**Setting 2: System Parse**		
	R	***P***	***F***	***R***	***P***	***F***
Baseline	75.1%	50.1%	60.1%	43.7%	30.7%	36.1%
Our Model	63.5%	**65.3%**	**64.4%**	**57.2%**	**55.7%**	**56.4%**

1) Results on Different Source of Data: We test our model on each source of the corpus, the results are shown in Table III.

TABLE III. EXPERIMENTSAL RESULTS ON EACH SOURCE OF DATA

	Setting1: Gold Parse			**Setting 2: System Parse**		
	R	***P***	***F***	***R***	***P***	***F***
NW	60.8%	62.4%	61.6%	57.5%	63.3%	60.3%
MZ	46.6%	57.0%	51.3%	44.9%	57.5%	50.4%
WB	65.4%	63.2%	64.3%	56.7%	55.9%	56.3%
BN	61.7%	67.6%	64.5%	52.0%	58.9%	55.2%
BC	66.0%	66.0%	66.0%	59.0%	52.7%	55.7%
TC	**76.5%**	**72.7%**	**74.6%**	**65.7%**	58.9%	**62.1%**

Results show that the performance differs greatly in different source of data. Considering that our approach models contextual information of a ZP forward from the beginning and backward from the end to current word. We argue that the small size of a sentence can avoid a potential forget of the "*history*" and the little number of ZP is benefit of capturing the contextual information continuously. We statics the average sentence length and average ZP number in each dataset, as shown in Table IV.

TABLE IV. STATISTICS OF EACH SOURCE OF DATA

	NW	**MZ**	**WB**	**BN**	**BC**	**TC**
AvgLen	25.9	35.3	22.4	30.9	16.4	**11.2**
NumZP	0.6	1.5	0.9	1.1	0.7	0.7

a. **AvgLen** means the average sentence length including the punctuations;

b. **NumZP** means the average zero pronoun number in one sentence.

Compared Table III with Table IV, we can observe that the results are basically consistent. As the average length of the TC is smallest among the six sources, and the average ZP number of it is relatively smaller, our model performs best in this source.

D. Error Analysis

To better evaluate our proposed approach, we perform a qualitative analysis on the errors obtained under Setting 1. Both the sentences in the corpus and the results obtained by our model are Chinese. The following examples are translated for non-Chinese readers.

We first analyze results that the detected ZP is inconsistent with the dataset.

In the first case, a different place for the ZP is predicted by our model. Within ZP resolution in the predicted place, the semantics is consistent with the standard sentence.

TABLE V. ERROR EXAMPLE 1

Corpus	After the advertisement, *pro* will broadcast international news for you.
Our Model	*pro*, after the advertisement, will broadcast international news for you.

From the example in Table V, although different with the corpus, the prediction made by our model is correct either. The adverbial modifiers are used flexibly in Chinese, which can be placed in the beginning or the end of the sentence, even between the subject and the predicate. Furthermore, we analyze the raw output of the network, the value for the annotated place in the corpus is the second top, which is 10 times larger than the third one.

In the second case, some new ZPs are detected by our model, which are not annotated in the corpus.

TABLE VI. ERROR EXAMPLE 2

Corpus	Making living beyond the hometown only last for two years, Wu Jian-le ... ...
Our model	*pro* has made living beyond the hometown only last for two years, Wu Jian-le ... ...

From the example in Table VI, the beginning of the sentence can be regarded as a ZP and use the subject of the second sub sentence (Wu Jian-le) to resolve it. From two kinds of cases above, we think that the ZP position is ambiguous. Thus, it is not suitable to set an absolute standard for ZP identification. We argue that it is better to determine the ZP position with a certain probability.

Finally, we check some errors from our model. A ZP after an object, which is the subject of the following clause, is hard to be recognized.

TABLE VII. ERROR EXAMPLE 3

Corpus	... ... has asked related department *pro* take some further investigate ... ...

From the example in Table VII, we can see that semantic information of both forward and backward directions are complete till the word “take”. Thus, our model predicts that there is no ZP. Although this prediction is incorrect, we think that our approach can model the semantic information to some extent.

E. Summary

In this section, we can see that under Setting 1 and Setting 2, our model outperforms the-stat-of-art method by 4.3% and 20.3% in terms of overall F-score respectively on OntoNotes 5.0 corpus. And we make some deeply analysis of the results.

● As average sentence length and average ZP number varies in each source of corpus, our model shows different performance and obtains the best performance in the Telephone Conversation corpus. Due to the smallest average sentence length (11.2) and smaller average ZP number (0.7) among six source datasets, our model easily obtains the contextual knowledge.

● We analyze some typical errors from our model. We discover that our model detects some ZPs are inconsistent with the dataset, but they make sense in some extent. We argue that it is better to determine the ZP position with a certain probability, rather than an absolute value. And we discover that our model is hard to recognize a ZP after an object which is the subject of the following clause. This is because from both forward and backward directions the semantic information is complete.

VI. Conclusions and Future Work

In this paper, we employ a stacked bidirectional LSTM to identity Chinese ZP. The main idea of our approach is investigating a leaning algorithm which can automatically extract the features from the raw text without feature engineering, and capably of modeling contextual semantic information to boost the performance. The experimental results on the OntoNotes 5.0 corpus show that our model significantly outperforms the state-of-the-art method.

In future, we plan to do some further researches:

1. employ an attention mechanism to model the long distance dependent contextual information.

2. develop a model for Chinese ZP resolution combined with our current approach.

Acknowledgment

The authors would like to thank two anonymous reviewers for giving the affirmation of the work and also acknowledge Katie and Denise from Linguistic Data Consortium (LDC) for their help of getting the OntoNotes 5.0 corpus.

References

[1] C. Chen and V. Ng, “Chinese zero pronoun resolution: A joint unsupervised discourse-aware model rivaling state-of-theart resolvers,” ACL 2015, Volume 2: Short Papers, 2015, pp. 320–326.

[2] C. Chen and V. Ng, “Chinese zero pronoun resolution with deep neural networks,” ACL 2016, Volume 1: Long Papers, pp778–788

[3] T. Liu, Y. Cui, Q. Yin, S. Wang, W. Zhang, and G. Hu, “Generating and exploiting large-scale pseudo training data for zero pronoun resolution,” CoRR, vol. abs/1606.01603, 2016.

[4] Q. Yin, W. Zhang, Y. Zhang, T. Liu: A deep neural network for chinese zero pronoun resolution. CoRR arXiv:1604.05800v2 (2016)

[5] S. Zhao and H. T. Ng, “Identification and resolution of chinese zero pronouns: A machine learning approach,” EMNLPCoNLL 2007, pp. 541–550.

[6] F. Kong and G. Zhou, “A tree kernel-based unified framework for chinese zero anaphora resolution,” EMNLP 2010, pp. 882–891.

[7] Y. Yang and N. Xue, “Chasing the ghost: recovering empty categories in the chinese treebank,” in COLING 2010,pp.1382–1390.

[8] Y.-J. Kim, “Subject/object drop in the acquisition of korean: A cross-linguistic comparison,” Journal of East Asian Linguistics, vol. 9, no. 4, pp. 325–351, 2000.

[9] Y. Bengio, “Learning deep architectures for AI,” Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1–127, 2009.

[10] C. Chen and V. Ng, “Chinese zero pronoun resolution: Some recent advances,” in EMNLP 2013, pp. 1360–1365.

[11] L. Wang, Z. Tu, X. Zhang, H. Li, A. Way, and Q. Liu, “A novel approach to dropped pronoun translation,” in NAACL HLT 2016, pp. 983–993.

[12] A. Graves, A. Mohamed, and G. E. Hinton, “Speech recognition with deep recurrent neural networks,” ICASSP 2013, pp. 6645–6649.

[13] F. A. Gers, J. Schmidhuber, and F. A. Cummins, “Learning to forget: Continual prediction with LSTM,” Neural Computation vol. 12, no. 10, pp. 2451–2471, 2000.

[14] A. Graves, Supervised Sequence Labelling with Recurrent Neural Networks, ser. Studies in Computational Intelligence. Springer, 2012, vol. 385.

[15] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, and et al., “Theano: new features and speed improvements,” CoRR, vol. abs/1211.5590, 2012.

[16] R. Al-Rfou, G. Alain, A. Almahairi, and et al., “Theano: A python framework for fast computation of mathematical expressions,” CoRR, vol. abs/1605.02688, 2016.

[17] F. Chollet, “Keras,” https://github.com/fchollet/keras, 2015.

[18] T.Tieleman and G.Hinton, “Lecture 6.5—rmsprop: Divide the gradient by a running average of its recent magnitude,” Tech. Rep., 2012.

[19] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” LREC 2010, pp. 45–50.

[20] https://dumps.wikimedia.org/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2