End-to-End Neural Context Reconstruction in Chinese Dialogue

Wei Yang,[1,2] **Qiao Rui,**[2] **Haocheng Qin,**[2]
Amy Sun,[3] **Luchen Tan,**[2] **Kun Xiong,**[2] **Ming Li**[1,2]
[1] David R. Cheriton School of Computer Science, University of Waterloo
[2] RSVP.ai [3] Huawei Inc

Abstract

We tackle the problem of context reconstruction in Chinese dialogue, where the task is to replace pronouns, zero pronouns, and other referring expressions with their referent nouns so that sentences can be processed in isolation without context. Following a standard decomposition of the context reconstruction task into referring expression detection and coreference resolution, we propose a novel end-to-end architecture for separately and jointly accomplishing this task. Key features of this model include POS and position encoding using CNNs and a novel pronoun masking mechanism. One perennial problem in building such models is the paucity of training data, which we address by augmenting previously-proposed methods to generate a large amount of realistic training data. The combination of more data and better models yields accuracy higher than the state-of-the-art method in coreference resolution and end-to-end context reconstruction.

1 Introduction

The chatbot is claimed to become a platform for the next generation of the human-computer interface. Recent researches on open-domain chatting systems (Lowe et al., 2017; Mei et al., 2015), open-domain question answering systems (Minaee and Liu, 2017; Chen et al., 2017) have shown promising results on single-round conversations. Meanwhile, most of these systems require the input question to be syntactically and semantically complete sentences. However, due to the language nature of humans, facing more than one round of conversation, we need to tackle the problem of contextual relationship where coreference and ellipsis occur frequently in dialogues leaving the sentence incomplete. The goal of context reconstruction in dialogues is to load context information from a multi-round dialogue, and remove the dependency on the previous contexts in the sentences, so that each sentence have complete and independent semantic meanings, so are answerable and processible by down-stream dialogue or question answering systems.

In this paper, we addressed the context reconstruction problem, which includes referring expression detection and coreference resolution in the dialogue domain. We present our part-of-speech (POS) tagging based deep neural network, including both the step-by-step models and the end-to-end model, for the detections and resolutions of coreference and ellipsis. Our coreference and ellipsis detection model reasons over the input sequence to detect the positions of coreference and ellipsis in the sentence. Our resolution model ranks the candidate entities with the input sentence where coreference and/or ellipsis are annotated. We also present an end-to-end detection-resolution network which consumes only the non-annotated input sentence and candidate entities. Our models utilize both the syntactic and semantic information by employing word embedding, convolution layers, and Long-short-term-memory (LSTM) units. Due to the lack of large well-annotated data, in this paper, we proposed a novel approach to construct annotated data in dialogue domain.

We summarize our contribution in this paper with three points: 1) We formulate the problem definition of context reconstruction in dialogue into one detection problem and one ranking problem and present the difference between it and traditional tasks such as pronoun and zero pronoun detection and mention candidate selection; 2) We present the analysis of the application of deep neural work for contextual resolution in dialogue, including both step-by-step and end-to-end approaches; 3) We propose a way to effectively construct a huge amount of silver data for the con-

Proceedings of the 1st Workshop on NLP for Conversational AI, pages 68–76
Florence, Italy, August 1, 2019.

text reconstruction task.

2 Related Work

There has been much classical or linguistic theoretical work on coreference resolution in texts. Coreference resolution is mainly concerned with two tasks, referring expressions detection, and mention candidate ranking. Referring expressions detection can be further divided into two subtasks: 1). find all words that do not have real meaning and refer to other mentions (他/he，她/she，它/it，这/this，那/that,...). We use the term 'pronoun' to represent these words without losing preciseness of linguistic definition in this paper. 2). find all zero pronouns. A close task to the first subtask of referring expressions detection is coreference detection, which is to identify noun phrases and pronouns that are referring to the same entities. Haghighi and Klein (2010) proposed an unsupervised generative approach for text coreference detections. Uryupina and Moschitti (2013) proposed a rule-based approach which employed parse trees and SVM. Peng et al. (2015) improved the performance of mention detections by applying a binary classifier on the feature set.

Similarly, there has been much previous work in mention candidate ranking using deep neural network. In recent years, applying deep neural networks on the task has reached great success. Clark and Manning (2016) applied reinforcement learning on mention-ranking coreference resolution. Lee et al. (2017) presented an end-to-end coreference resolution model which reasons over all the anteceding spans. Lee et al. (2018) presented a high-order coreference resolution. These approaches do not generalize to dialogue for the reason that 1) these approaches require a rich amount of well-annotated contextual data, 2) dialogue is short and has ambiguous syntactic structures which are difficult to handcraft rules, and 3) the resolution module should distinguish wrong detection results so that the systems have a higher fault tolerance on the detection module. However, most existed work simply assumes a golden detection label and perform lots of feature engineering based on that.

Although there is a series of related work that can contribute to coreference resolution in Chinese dialogue, there are many common restrictions when transferring them into a practical product: 1). the limited data source in a general domain; 2). most work concentrates on general coreference. Few of them focus on pronoun or zero pronoun resolution, which is the vital step for dialogue NLU; 3). no work known to us compares traditional feature-based methods and neural network based models on an end-to-end system for coreference resolution in Chinese dialogue.

Context (**c**): 打雷 了 怎么 发 短信 安慰 女朋友？
(How to send texts to comfort girlfriend when it thunders?)
Text (**q**): 打雷 时 还给 她 发？
(Send to her even when it thunders?)
Text (**q**) after detection: 打雷 时 还给 她 发 ϕ？
Text (**q**) after resolution: 打雷 时 还给 **女朋友** 发 **短信**？
(Send **texts** to **your girlfriend** even when it thunders?)

Figure 1: Example of context reconstruction

3 Our Approach

Figure 1 provides a running example of our context reconstruction approach. We assume an input utterance **q** whose context we are trying to reconstruct with respect to some other context utterance **c**. In the chat context, **c** would come from previous utterances in the dialogue. In a benchmark dataset, we locate the context using the first sentence where the co-referred mention appears. We assume that **q** and **c** have already been tokenized. Our approach breaks the context reconstruction problem into two subtasks: detection and resolution.

Detection is formulated as a sequence labeling task that tries to identify referring expressions that need to be resolved and to recover zero pronouns. In our running example, 她 (her) is identified as such, as well as a zero pronoun ϕ (an elided object). Resolution is formulated as a ranking task. For each "slot" that needs to be resolved (她 and ϕ in the example above), our model provides a ranking of $(\mathbf{c}, \mathbf{q}, m)$ triplets, where $m \in \{m_1, \ldots, m_k\}$, the candidates for resolution. Candidates are selected from noun phrases in the context **c**. At inference time, the candidate m with the highest score is selected as the replacement. If there are multiple slots to be resolved, our model proceeds from left to right incrementally. The final output of the model is shown in the last line of Figure 1. In this paper, we call our POS tagging based model as POSNet. The detection and ranking part is named POSNet-D and POSNet-R accordingly.

3.1 Detection

The detection subtask attempts to identify referring expressions that need to be resolved and to recover the position of zero pronouns. Note that not all referring expressions require resolution. For example, '这' (this) in '这个理由很有说服力' (This reason is convincing) requires no resolution, while '这' (this) in '这个不是我想要的' (this is not what I want) does. Detection is formulated as a sequence labeling task where the output labels $y \in [0, 1, 2]$. The label '1' indicates the boundary of a "slot" while the label '2' is assigned to expressions requiring resolution. Thus, in our running example, the input [PAD 打雷 时 还给 她 发 PAD] would be tagged with [0 0 0 1 2 1 0]. That is, the pronoun '她' is explicitly tagged, together with its left and right boundaries; consecutive '1' tags indicates a zero pronoun.

In our detection model, the (padded) sentence and POS tagging encoding layer consists of the following components: First, we apply 200-dimensional embedding layer (Mikolov et al., 2013) to $\mathbf{s}$ and a 20-dimensional embedding layer to $\mathbf{t}$. Let $\mathbf{s} = \{s_1, \ldots, s_m\}$ and $\mathbf{t} = \{t_1, \ldots, t_m\}$ be the embedded representations. To leverage to position information which is important in this task, we also include the position embeddings suggested by Gehring et al. (2017) in the model with the same size as the word embedding, denoted as $\mathbf{p} = (p_1, \ldots, p_m)$. The word embeddings and POS embeddings are incorporated together by summing and then concatenated with the position embedding as the combined input: $\mathbf{w} = \{w_1, \ldots, w_m\}, w_i = [s_i + p_i, t_i]$.

Inspired by the recent success of convolutional models for various NLP tasks (Kim, 2014), we apply a stack of 5 convolution layers followed by a global max pooling layer on top of the word and POS tagging encodings to extract underlying patterns in the sentence. We use gated linear units (GLU) (Dauphin et al., 2016) as the activation function, and we included residual connections to reduce training difficulty (He et al., 2016). After the encoding the input using convolutional layers with residual connections, we apply LSTM as the decoder to generate the sequential predictions for the location of referring expressions as $\{d_1, \ldots, d_n\}$. To train this model, we apply categorical cross entropy loss $\mathcal{L}_{seq}$ over a text sequence:

$$\mathcal{L}_{seq} = -\frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_{\text{class}}} y_{ij} \log(d_{ij})$$

3.2 Resolution

The output of the detection model is a list of "slots" that require resolution, which could either be a referring expression or a zero pronoun. In the resolution task, for each slot, the model finds the most appropriate replacement to best reconstruct the context. This is formulated as a ranking problem over $(\mathbf{c}, \mathbf{q}, m)$ triplets, where $m \in \{m_1, \ldots, m_k\}$ are the candidate mentions for resolution. In our running example, there are two slots to be resolved (她 and ϕ); at inference time, our model selects the highest scoring m for each slot, proceeding from left to right.

The input to the model comprises a sentence, its corresponding POS tags, a known pronoun or zero pronoun slot, and a candidate mention. Then, we concatenate word embeddings and POS tagging embedding as the input of mentions and encode it using multilayer perceptron. To enrich the semantic information of the mention candidate, we find the context sentence that contains this mention as another input. Usually this context is the sentence exactly before the query sentence in dialogues. Then we encode the query and context in the same way described in Section 3.1. We did not add attention mechanism, as the interaction method as described by Yin et al. (2018b) to our model because we did not see significant improvement with preliminary experiments. To train the mention candidate ranking model, we apply hinge loss to maximize the margin between a positive sample and a negative sample as below:

$$\mathcal{L}_{\text{hinge}} = \max\{0, \\ \delta + \mathcal{F}(\mathbf{w}_q, \mathbf{w}_c, \mathbf{m}^-) - \mathcal{F}(\mathbf{w}_q, \mathbf{w}_c, \mathbf{m}^+)\}$$

where $\mathcal{F}(\cdot)$ is the ranking model. $\mathbf{w}_q$ and $\mathbf{w}_c$ are the input with words, POS tagging and position embeddings of query and context. $\mathbf{m}^-$ and $\mathbf{m}^+$ are the positive and negative mention embedding including the POS tagging embedding. δ is a hyper-parameter and we set $\delta = 1$ in our experiments.

3.3 End-to-End Reconstruction

When combining the detection and ranking modules, we propose a masking structure to add a

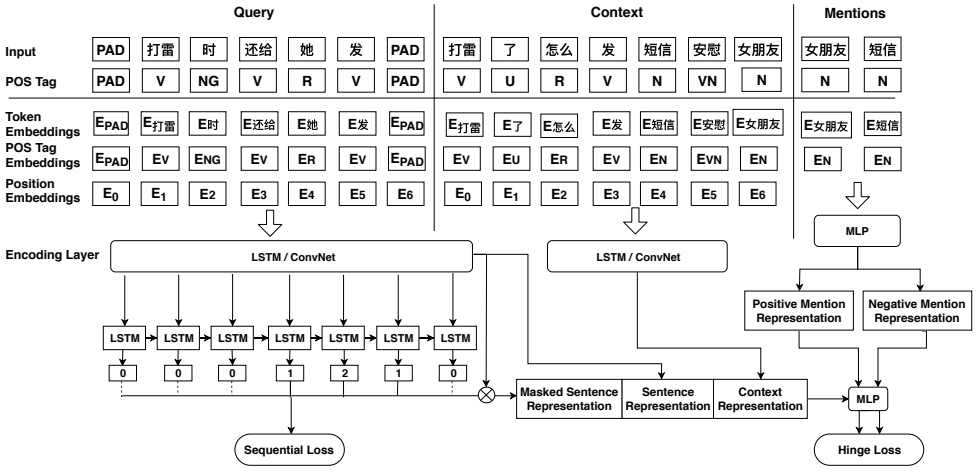

Figure 2: Architecture of the neural end-to-end model for coreference resolution in Chinese dialogue

masked sentence representation layer in the joint model. The mask vector is from the sequential prediction of the detection module, and we apply it back to encoded sentence matrix to highlight the words near the pronoun or zero pronoun slot to get the masked sentence representation $\mathbf{v_{ms}}$:

$$\mathbf{v_{ms}} = \text{Pooling}(M_s \mathbf{v_m})$$

where $\mathbf{v_m}$ is the binary mask vector and M_s is the encoded sentence representation matrix. A max pooling function is applied to project the masked sentence matrix into a vector. Through this way we try to force the model to selection mention candidate that is mostly likely to co-occur near a pronoun or zero pronoun. These words are usually verbs (e.g. love, publish) but seldom prepositions (e.g. through) or adjectives (e.g. wonderful). Based on the above two individual models, we combine the learnt (masked) sentence representation and the mention representation and build the end-to-end context reconstruction model (or joint model), where the detection and resolution models are trained jointly. The overall framework is shown in Figure 2.

To train this model, we combine the hinge loss $\mathcal{L}_{\text{hinge}}$ and the sequential loss $\mathcal{L}_{\text{seq}}$ mentioned above. The two losses are aggregated by a hyper-parameter λ for the trade-off. Finally, we add a regularization term to the target function to reduce overfitting. The final loss can be written as follows:

$$\mathcal{L} = \mathcal{L}_{\text{hinge}} + \lambda \cdot \mathcal{L}_{\text{seq}} + \mu \cdot ||W||$$

where λ and μ are hyper-parameters, and $||W||$ is the regularization term over all weights in the model. When integrating the POSNet-R with POSNet-D, we find that sometimes POSNet-D predicts a word in a sentence to be a reference when it is not. This requires our POSNet-R to have the ability to predict that nothing fits for a wrong slot detection. To achieve this, we create a special mention candidate UNK, representing the null string. At inference time we can input UNK along with other candidates NPs to POSNet-R. If UNK token has the highest score, that means nothing should be fit into the reference slot. We trained POSNet-R again with the aforementioned modifications on the same training data set. Thus, we modify the hinge loss as below:

$$\begin{aligned}\mathcal{L}_{\text{hinge}} = &\max\{0, \delta + \mathcal{F}(\mathbf{w}_c, \mathbf{w}_q, \mathbf{m}^-) \\ &- \mathcal{F}(\mathbf{w}_c, \mathbf{w}_q, \mathbf{m}_0)\} \\ &+ \max\{0, \delta + \mathcal{F}(\mathbf{w}_c, \mathbf{w}_q, \mathbf{m}_0) \\ &- \mathcal{F}(\mathbf{w}_c, \mathbf{w}_q, \mathbf{m}^+)\}\end{aligned}$$

Where $\mathbf{m}_0$ represents the embedding for UNK.

Data	Docs	Sents	ZP
CONLL2012$_{Train}$	1,394	36,487	12,111
CONLL2012$_{Test}$	172	6,083	1,713
OntoNote$_{BC}$	-	2,800	1,400
OntoNot$_{TC}$	-	1,628	814

Table 1: Statistics of the CONLL2012 and the OntoNote datasets

4 Experimental Setup

4.1 Dataset

We conduct all of our experiments on Chinese datasets. Note all of our models used in this pa-

Type	Neg	ZP	Pronoun	Total
NP	1M	800 000	1 200 000	3 000 000
Location	1M	200 000	750 000	1 950 000
Person	1M	200 000	750 000	1 950 000
Time	1M	990 000	601 000	1 700 000

Table 2: Statistics of the generated CQA dataset

per are language-independent. We have evaluated our models on three datasets. The statistics of all datasets is shown in Table 1 and Table 2.

- CONLL2012: To get a fair comparison with the previous methods, we applied POSNet-R to the zero pronoun resolution task on the CONLL2012 benchmark dataset following Yin et al. (2018a) and Yin et al. (2018b)'s processing methods. Note this is the dataset annotated with the coreference of zero pronouns in a general domain and this task assumes the pre-known location of zero pronouns so we apply POSNet-R as a comparison.

- OntoNote (BC/TC): Since there is no known end-to-end evaluation benchmark for Chinese context reconstruction, we extracted data from the BC (broadcast conversation) and TC (telephone conversation) subsets from OnotoNote 5.0 corpus (which is the same source of CONLL2012) and build the end-to-end training and evaluation dataset for zero pronoun resolution. We apply basic cleaning on the corpus such as removing the cataphoric reference and filling multiple coreferences in one sentence. For each sentence with a zero pronoun, we sample one negative candidate from the last sentence and use this sentence as a context sentence.

- CQA: Since CONLL2012 and OntoNote are either too small to evaluate the performance of neural network or too domain-specific to provide a satiated training and evaluation on a general domain, we collected and built new training and testing set from Chinese CQA (community question answering website) websites including BaiduZhidao[1], SosoWenwen[2], which contains over 300,000,000 QA pairs. We generated *time*, *location*, *people* and *noun phrase* examples. Each subset is divided into the training data and the testing data at the ratio of 9:1. We use this generated data to mimic the coreference in the real data and we will show this generated data contributes to both general evaluation and external assistance to a specific domain.

[1]https://zhidao.baidu.com/
[2]https://wenwen.sogou.com/

4.2 Dataset Generation

Contextual resolution on dialogue corpus requires large-scale and annotated training data. Obtaining such a data set is the key to this problem. We introduce our three-phases data generation method as follows: data collection, keywords detection, and data splitting.

Data Collection: Sentences in dialogues have the features of being short and containing only one or two entities. Corpus from CQA websites fit our purpose perfectly since 1). these questions and answers tend to be short and precise; 2). large user groups provide a huge corpus of data; 3). these single round question-answering dialogues share some language features with chatting dialogues. Initially, QA pairs from the internet are collected. These are our *raw data*. These raw data are mostly precise, complete, short, and independent sentences and contain no coreferences to the context.

Keyword Detection: First of all, we detect and label words that refer to *time*, *location*, *people* or *noun phrases*. We parse questions using the Parser (Roger Levy, 2003) to generate syntax trees annotated with POS taggings. The POS taggings provide syntactic information that helps guide the data generation rules. Then, we use the Stanford named entity recognizer (Finkel et al., 2005) to tag tokens that refer to *time*, *location* or *people* entities, named *marked words*.

Data Splitting: Our goal is to transform short sentences from dialogues into positive examples of coreference and ellipsis. The main challenge in generating those is to identify segments that can be omitted or replaced with a pronoun so that the resulting sentence is both grammatical and natural. Our method splits complete sentences into sentences that contain pronoun or zero pronoun according to the self-defined syntactic pattern: 1) Pronoun samples: Since pronouns actually refer to an entity from the context, we can reverse the process and create coreference cases by replacing entities with pronouns in sentences. It is feasible also

because for a certain entity type (e.g. time), the corresponding pronouns are limited. 2) Zero pronoun samples: For the same reason as above, the process of understanding zero pronouns could be reversed. We can create ellipsis cases by omitting entities in sentences. Therefore, we create ellipsis cases by deleting the marked words in the sentence directly. 3) Negative samples: There are two types of negative samples in this problem. The first type is a sentence without generated pronoun or zero pronoun. In order to provide competitive samples for training, negative examples are randomly sampled out of the whole CQA corpus. In addition, a number of complete sentences that contain pronouns and zero pronouns already are added. It could enhance our model's ability to distinguish real coreference and "fake" coreference. The second negative samples are the mention candidates that are not referred to. We randomly sample mentions from the same session or document to make the negative samples challenging.

4.3 Model Training

We use Jieba[3], a Chinese word segmentation tool to segment a sentence into a sequence of words. The Chinese word embeddings are pre-trained using skip-gram model (Mikolov et al., 2013) on the raw CQA corpus. The LSTM-encoder and LSTM-decoder in all of our models have a state size of 512. The convolution layers have 512 filters with width 3. The models are trained by the Adam optimization algorithm (Kingma and Ba, 2014) with a learning rate of 3×10^{-4}. Vocabulary size is truncated by selecting the most frequent 200,000 tokens. λ is set to 20 and μ is set to 0.01 in all of our experiments.

5 Results

5.1 Detection

Although we model referring expression detection as a sequence labeling task, we assume there is at most one pronoun or zero pronoun in a sentence. So we report sentence-level precision, recall, and F_1 scores for evaluation in coreference resolution task in dialogue. Note we can run this detection algorithm iteratively after one round of context reconstruction if the sentence contains multiple pronouns or zero pronouns in practical application. The experimental results on CQA dataset are shown in Table 3.

[3]https://github.com/fxsjy/jieba

Data	Pre.	Rec.	F_1
Name phrase	92.7	96.9	94.8
Location	95.3	95.7	95.5
Person	92.9	97.5	95.1
Time	91.1	95.7	93.3
Average	93.0	96.5	94.7

Table 3: Results of POSNet-D for referring expression detection on CQA dataset

Model	P@1	P@2	P@3
Bigram	22.8	37.1	48.2
Yin et al. (2018b)	68.1	87.3	89.5
Yin et al. (2018a)	68.3	**87.7**	89.7
POSNet-R	**69.1**	85.2	**91.2**

Table 4: Results of mention candidate ranking on the CQA dataset

According to Table 3, the high F1 scores indicate the strong ability of POSNet-D to distinguish positive examples and negative examples. The slightly higher recall rate than precision indicates the model tends to treat potentially words as positive and retrieve more potentially positive candidates, which meets our requirement to provide more candidates for ranking in this detection step properly. Note that from Table 3, we can also find the accuracy on location and people subsets is higher than NP and time. This is because there are more ellipse detection cases in NP and time subsets, which bring a challenge to our model and baseline method by causing more false negatives.

5.2 Resolution

We test mention candidate ranking on two datasets: CQA and CONLL2012. For each sentence in the test set, we feed it into the model together with the correct mentions and nine randomly sampled mentions. The model outputs the ranking scores for all 10 mentions and we choose the one with the highest score as the model's prediction. Under this setting, a naive model that outputs random scores should result in an overall top 1 accuracy close to 10%. The overall performance is shown in Table 4. Bigram in Table 4 is the baseline method that we select the candidate with the largest co-occurrence frequency with the preceding and the following word as the prediction. Additionally, POSNet-R pretrained on the CQA dataset outperforms al baselines, which demonstrates the effectiveness of our generated data.

Model	F_1	P@1
POSNet-D+Yin et al. (2018a)	92.9	69.7
POSNet-D+Yin et al. (2018b)	92.9	69.9
POSNet	**95.4**	**71.7**

Table 5: Results of the end-to-end evaluation for coreference resolution on the CQA dataset

Model	F_1
Zhao and Ng (2007)	41.5
Chen and Ng (2016)	52.2
Yin et al. (2017)	54.9
Liu et al. (2016)	55.3
Yin et al. (2018b)	57.3
Yin et al. (2018a)	57.2
POSNet-R (raw)	52.1
POSNet-R (pretrained on CQA)	**58.1**

Table 6: Results of mention candidate ranking for zero pronouns on the CONLL2012 dataset

For the CONLL2012 dataset, the result is shown in Table 6. Following Yin et al. (2018b), we add the features from existing work on zero anaphora resolution into the fully connection layer. We try POSNet-R and find it performs close to the previous nerual network methods but cannot beat the Yin et al. (2018b)'s model. We think this is because our model needs more training data to learn an effective representation of the text and POS tagging so we pretrain our model on the whole CQA dataset. The result shows we can achieve the best performance on this benchmark.

5.3 End-to-end Evaluation

End-to-end model is tested on two datasets: the generated CQA and the extracted OntoNote. This model is trained with the original sentence as well as the correct NP and 9 sampled negative NPs. The output consists of two parts, the coreference and ellipsis detection of the sentence, and the ranking score of the mention candidate. The experiment results of the end-to-end evaluation on CQA and OntoNote datasets are shown in Table 5 and Table 7. Comparing the results of the joint model (Table 5) with the Table 3, we found that the end-to-end model has improvements on the F1 score. We find that it is because the precision score increases while the recall score drops a little. This result shows that involving candidate phrase information, the ability to detect the correct coreference and ellipsis is improved. Comparing to the joint model with the POSNet-R, we found that the top 1 accuracy is slightly improved, while top 2 and top 3 accuracies are dropped. The drops are expected as the position information of coreference and ellipsis are not given.

Test	Pretrain	Train	F_1	Accuracy
TC	CQA	BC	45.3	92.5
		-	10.4	66.8
	-	BC	18.3	72.5
BC	CQA	TC	36.1	84.5
		-	11.8	65.0
	-	TC	16.2	69.2

Table 7: Results of end-to-end zero pronoun resolution on OntoNote dataset

Since there is no known end-to-end Chinese context reconstruction model for the dialogue corpus, we compare POSNet with two step-by-step baselines: POSNet-D for the detection first, Yin et al. (2018a) and Yin et al. (2018b)'s methods for the ranking next. Comparing to the joint model with the baselines, we can see that step-by-step approach will cause serious cascade error if one step cannot perform well. In contrast, our model joint performs reasonably well considering the returned top 3 candidates. However, to better help the down-stream natural language understanding task, we should mainly aim at transforming a sentence extracted from the dialogue corpus to an independent sentence. So accuracy at top 1 is the most important evaluation metric.

We shows the results on OntoNote dataset in Table 7. From the result of these two small data sets we can see it is important to 1). learn a general knowledge by pretraining on a large corpus; 2). fine tune on a domain-specific dataset to get the downstream information such as common terms, common grammar, etc. In addition, by looking at Table 5 and 7 together, we can see that coreference detection, especially zero pronoun detection, is the bottleneck of the end-to-end context reconstruction system.

5.4 Ablation Study

We compare our model to the following ablated models: replacing the encoding layer with the BiLSTM layer, removing the UNK token candidate, removing word position embedding, and removing POS tagging from the input. The results are shown in Table 8. From Table 8 we find that

Model	P@1	P@2	P@3
POSNet	70.1	82.9	89.0
POSNet-LSTM	68.1	82.2	90.2
POSNet w/o UNK	67.4	81.2	86.2
POSNet w/o pos-embed	67.2	81.0	88.1
POSNet w/o POS input	61.8	71.4	73.7

Table 8: Alation study of the end-to-end contexual resolution on the CQA dataset

POSNet achieves better performance than the base POSNet model without UNK augmentation. We believe it is because 1) the UNK token helps enlarges the distance between the relevance of positive samples and negative samples. 2). it allows the mention candidate ranking model to identify the false positive of the detection model and replace it with a rejection token.

In addition, we try BiLSTM as the encoder as the comparison to the CNN based encoder in the experiments and we name it POSNet-LSTM. From the result, we can see BiLSTM gives weaker performance than ConvNet layers. We argue that this is because ConvNets layers are more sensitive to the distant and global dependency information in coreference while LSTM cares more about adjacent words. From the result of removing position embedding and the POS input, we can see that this task heavily relies on the understanding of the sentence syntactic structure. We believe there will be better ways to leverage this kind of information in a sentence.

6 Conclusion

In this paper, we systematically define the context reconstruction problem in dialogue domain and initiated a comprehensive study of this problem. We have demonstrated how to create training data to train both two step-by-step neural networks and an end-to-end deep neural network to tackle this problem. This study leads to many open research directions. Our work could be extended to wider contextual domains, including more conjunctive relations and more careful linguistic studies of conjunctive relations in conversations. Studies could go beyond context reconstruction and include semantics from conversation history. At the application level, neural context reconstruction can be easily integrated with an end-to-end question answering system (Yang et al., 2019) for a extrinsic evaluation.

References

Chen Chen and Vincent Ng. 2016. Chinese zero pronoun resolution with deep neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 778–788.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Kevin Clark and Christopher D Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2016. Language modeling with gated convolutional networks. *arXiv preprint arXiv:1612.08083*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. *arXiv preprint arXiv:1804.05392*.

Ting Liu, Yiming Cui, Qingyu Yin, Weinan Zhang, Shijin Wang, and Guoping Hu. 2016. Generating and exploiting large-scale pseudo training data for zero pronoun resolution. *arXiv preprint arXiv:1606.01603*.

Ryan Thomas Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65.

Hongyuan Mei, Mohit Bansal, and Matthew R Walter. 2015. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. *arXiv preprint arXiv:1509.00838*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Shervin Minaee and Zhu Liu. 2017. Automatic question-answering using a deep similarity neural network. *arXiv preprint arXiv:1708.01713*.

Haoruo Peng, Kai-Wei Chang, and Dan Roth. 2015. A joint framework for coreference resolution and mention head detection. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 12–21.

Christopher D. Manning Roger Levy. 2003. Is it harder to parse chinese, or the chinese treebank? pages 439–446. Association for Computational Linguistics.

Olga Uryupina and Alessandro Moschitti. 2013. Multilingual mention detection for coreference resolution. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 100–108.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.

Qingyu Yin, Yu Zhang, Weinan Zhang, and Ting Liu. 2017. Chinese zero pronoun resolution with deep memory network. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1309–1318.

Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018a. Deep reinforcement learning for chinese zero pronoun resolution. *arXiv preprint arXiv:1806.03711*.

Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018b. Zero pronoun resolution with attention-based neural network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 13–23.

Shanheng Zhao and Hwee Tou Ng. 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 (EMNLP-CoNLL)*.