Research on Coreference Resolution Based on Conditional Random Fields

Yujie Miao\textsuperscript{1,}\textsuperscript{*}, Xueqiang Lv\textsuperscript{1}, and Le Zhang\textsuperscript{1}


\textsuperscript{1}Beijing Information Science and Technology University, Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing, 100101, China

Abstract. In view of the phenomenon of noun coreference in Chinese, this paper proposes a deep learning mechanism based on Conditional Random Field (CRF) to study coreference resolution based on deep semantic information representation. The text is input into the vector of Biditive Encoder Representations from Transformers model. The self-attention mechanism is used to mine the hidden features at the context semantic level. Through the reasoning ability of CRF, the complex features are used for reasoning training, and the training results are scored and classified by softmax to complete the anaphora resolution task. The experimental results show that the performance of coreference resolution can be effectively improved by making full use of text feature representation.

1 Introduction

Coreference is a widespread phenomenon in natural language and an important expression, making coreference resolution a key task in the field of natural language processing. Coreference resolution can be regarded as a clustering or classification process, which mainly judges whether entity expressions point to the same entity or whether entities can be added to a set of entities. Coreference resolution is a key task in the field of natural language understanding. Accurate and unambiguous coreference resolution can promote the overall understanding of text semantics and plays an extremely important role as a basic supporting structure for natural language applications such as information extraction, machine translation, question answering system, etc.

After a long period of research on coreference resolution, some classical methods have emerged. For example, Pradheep summarized the previous work in 2005 and systematically divided coreference resolution methods into linguistic-based resolution methods and machine learning-based resolution methods for the first time [1]. In 2007, Lang Jun and others analyzed the technical route of coreference resolution in Chinese and English for more than 30 years, introduced in detail the international evaluation methods of coreference resolution, listed relevant resources and tools, and made a general prediction of the future development direction of coreference resolution [2]. In 2010, Vincent made a very systematic and detailed review of some important supervised learning methods in the past 15 years according to the steps of supervised learning [3], pointed out the shortcomings of coreference resolution methods based on supervised learning, and gave two solutions. With the in-depth research on coreference resolution, the focus of research is also more inclined to data-driven methods, such as methods based on supervised learning [4] and methods based
on unsupervised learning [5]. At present, a trend of coreference resolution is how to formally apply deep grammar, semantic knowledge and structural information to coreference resolution, that is, how to obtain deep semantic information.

In recent years, with the rise of neural networks, words can be expressed as vectors that transfer semantic dependencies [6]. Dependencies between words can be captured by structures such as cyclic neural networks. In addition, neural networks have excellent data fitting and classification capabilities. More and more scholars have begun to apply various neural network models to coreference resolution tasks. Typical works include: Clark et al. [7] adopted reinforcement learning and used its Reward Rescaling method to resolve coreference, and obtained the best known performance on CoNLL 2012 [8] Chinese test set. Wu et al. [9] proposed to use convolution neural network to represent the context of word embedding, then to represent the Mention, and to use classifier for final resolution. Lee et al. [10] proposed an end-to-end entity coreference resolution model based on neural network. The model uses bidirectional LSTM and Head-finding attention mechanism to represent phrases, and uses phrase sorting model to complete coreference resolution.

On the basis of previous work, Lee et al. [11] used the resolution process to iteratively update the expression of phrases, trimmed the candidate antecedent search space of the term to be resolved by bilinear attention, and obtained the best known performance on CoNLL 2012 English test set in combination with ELMo(Embedding from Language Model) [12]. Thanks to the excellent characterization ability of neural networks, the performance of coreference resolution model based on neural networks has been greatly superior to that of traditional models. With the emergence of neural networks and deep learning models, semantic representation has been greatly improved. Previous works mostly use Bi-LSTM (Bi-directional Long Short Term Memory) or CNN (Convolutional Neural Networks) to encode sentences [13, 14], but the disadvantage is that Bi-LSTM is difficult to parallelize and CNN has insufficient ability to capture global information. Recently, some pre-trained language models, such as ELMo and BERT (Bidirectional Encoder Representations from Transformers) [15], have been used in more and more NLP tasks. BERT pre-training model shows significant advantages in text data learning [16, 17], which not only reduces training time, but also effectively improves network performance. In this paper, the conditional random field model is introduced for sequence labeling, which is supervised and applied to a large number of data. It fully represents the context relationship and is convenient to express long-distance correlation. It can well solve parallel computing and context information representation.

2 Task model

The model diagram is shown in the following figure. First, Bert model is used to embed the text, then the full text is used to represent the context features, then CRF is used to better represent the long-distance information, and finally the feature vectors are scored.
2.1 BERT embedding

BERT model is a pre-training language representation model developed by Google. The pre-training model uses two-way transform structure and innovative MLM (Masked Language Model) to train the language model, which has been pre-trained on a large number of tagged text corpus. BERT model converts each word in the text into a one-dimensional vector by querying the word vector table as model input. The model output is the vector representation after inputting the corresponding fused full-text semantic information of each word. However, the model input includes two other parts besides the word vector: the first is the text vector. This vector automatically learns its value during model training. Global semantic information used to represent text. And fused with the semantic information of each word; The second is the position vector. Different positions in the text will lead to different semantic information representations. Therefore, BERT model will add a different vector to distinguish words in different positions. Finally, the word vector, text vector and position vector are added, and the sum is used as the input of BERT model.

2.2 Self-attention Layer

The second layer is the self-attention layer. Self-attention does not depend on the order between words. By calculating the similarity between words to mine information, information loss can be avoided. When Bi-LSTM processes sequences, if it does not calculate the results of the previous time, it cannot calculate the results of the next time, thus causing parallel computation. Self-attention can fully consider the semantic and grammatical relations between different words in sentences. The word vector obtained after calculation takes into account the connection and details between contexts. At the same time, it can perform parallel operations well and greatly improve the computational efficiency. Therefore, the second layer takes self-attention. The Eq. 1 of the Self-attention layer is as follows:

$$Attention(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$ (1)
Q (Query) refers to matching with other items, K (Key) refers to the matched items, and V (Value) is the extracted information. Its meaning is to use attention weight to carry out weighted linear combination on the mathematically expressed word vectors, so that each word vector contains information of all word vectors in the current sentence.

2.3 CRF

CRF [18] is a statistical model. It is developed from Hidden Markov Model (HMM) and Maximum Entropy (ME), and has the advantages of the two models at the same time. CRF can use complex feature functions to judge specific information. In the training process, CRF can make full use of text context information specified by feature functions to adjust parameters, thus applying context information in the process of reasoning and word segmentation. CRF is especially good at dealing with problems in the field of natural language understanding. In the process of feature extraction, conditional random fields can consider farther-away full-text features, accommodate a lot of context information, and flexibly design features. At the same time, CRF not only normalizes a certain word, but also normalizes the features of all words in the full text. Therefore, CRF model is widely used in various scenarios. The main Eq. 2 are as follows:

\[
p(l|s) = \frac{\exp[\text{score}(l|s)]}{\sum_{l'} \exp[\text{score}(l'|s)]}
\]  

3 Experiment and result analysis

3.1 Experimental data

The corpus comes from 10,000 news data crawled from major energy websites. The corpus is characterized by rich emotion, rigorous content, smooth writing, no rigidity in the writing form of news, and no fixed way of feature expression.

3.2 Experimental settings

For BERT model, Google provides two pre-training models, one is BERT-Base and the other is BERT-Large. The two models are the same except for some parameters. In this experiment, BERT-Base model is adopted due to the small scale of data. We set the maximum sequence length to 80. The training is divided into two stages, the first stage is pre-training. In this article, 256 sentences are taken as a batch, each sentence has a maximum of 512 tokens, and drop_out is set to 0.1. The second stage is fine tuning. In the fine-tuning phase, the super-parameters of most models are similar to those of pre-training, in which the learning rate is set to 5E-5. In the model training, the Adam algorithm is used to update the parameters and set the minimum loss learning rate to 0.001 and dropout to 0.5.

3.3 Result and result analysis

In order to show the advantages of the model, we carried out comparative tests and compared Bert model with the most benchmark model. The comparison models are as follows: 1) CNN model; 2) RNN model; 3) Bi-LSTM model; 4) BERT model; 5) BERT + self-attention model; 6) BERT + CRF model; 7) This model. The evaluation criteria are Precision, Recall and F value. The results are shown in Table 1 below. Through the comparison results, the overall F
value of CNN, RNN and BERT models is lower than that of BERT and Bi-LSTM, which is due to the relatively weak performance of CNN and RNN in mining deep semantic volumes of data compared with BERT and Bi-LSTM. The latter can mine semantic information from the hidden depth, capture the complex distribution of data and learn high-level features, thus laying a good foundation for improving the accuracy of coreference resolution. At the same time, with the addition of self-attention mechanism and CRF model, the coreference resolution performance has been further improved. Therefore, the accuracy of the model proposed in this paper is higher than that of the comparison model, which shows that the model in this experiment can accurately extract the relevant features of words and has strong coreference resolution performance.

### Table 1. Model performances.

| Model               | Precision | Recall | F-Measure |
|---------------------|-----------|--------|-----------|
| CNN                 | 36.85     | 49.03  | 43.70     |
| RNN                 | 34.12     | 46.98  | 41.25     |
| Bi-LSTM             | 42.48     | 58.36  | 52.91     |
| BERT                | 48.77     | 67.82  | 58.39     |
| BERT+self-attention | 52.13     | 69.31  | 61.78     |
| BERT+CRF            | 49.62     | 63.24  | 58.03     |
| This paper          | **56.08** | **72.59** | **66.46** |

### 4 Summary

This paper proposes a model of coreference resolution based on conditional random fields. The model starts from the semantic information of the text, BERT model is used to model and represent all the words in the text, and conditional random fields are used to consider the features of the full text at a longer distance. However, we can see from the results that the model still has some problems, such as the failure to represent the information in the text with good features, resulting in poor results. Therefore, in the future, we will focus on how to integrate prior knowledge into the research of coreference resolution.

### References

1. Elango P. Coreference resolution: A survey[DB/OL].[2012-03-12].http://ncce.inaoep.mx/~villasen/index_archivos/cursoTATII/EntidadesNombradas/Elango-SurveyCoreferenceResolution.pdf.
2. Lang Jun, Qin Bing, Liu Ting, et al. A review of the research on the resolution of textual total reference [J]. Journal of Chinese Language and Computing, 2007, 17 (4): 227-253.
3. Zhang R, Santos C N, Yasunaga M, et al. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. arXiv preprint arXiv:1805.04893, 2018.
4. Ng V. Supervised noun phrase coreference research: The first fifteen years[C]//Proceedings of the 48th annual meeting of the association for computational linguistics. Uppsala: Association for Computational Linguistics, 2010:1396-1411.
5. Ng V. Unsupervised models for coreference resolution[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Honolulu: ACL, 2008:640-649.
6. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems. Lake Tahoe: NIPS, 2013: 3111-3119.

7. Clark K, Manning C D. Deep Reinforcement Learning for Mention-Ranking Coreference Models[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: EMNLP, 2016: 2256-2262.

8. Pradhan S, Moschitti A, Xue N, et al. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes[C]//Joint Conference on EMNLP and CoNLL-Shared Task. Jeju Island: ACL, 2012: 1-40.

9. Wu J L, Ma W Y. A deep learning framework for coreference resolution based on convolutional neural network[C]//2017 IEEE 11th International Conference on Semantic Computing (ICSC). San Diego: IEEE, 2017: 61-64.

10. Lee K, He L, Lewis M, et al. End-to-end Neural Coreference Resolution[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017: 188-197.

11. Lee K, He L, Zettlemoyer L. Higher-Order Coreference Resolution with Coarse-to-Fine Inference[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans: ACL, 2018, 2: 687-692.

12. Peters M E , Neumann M , Iyyer M , . Deep contextualized word representations[J]. 2018.

13. Biao Z, Jinsong S, Deyi X, et al. Shallow convolutional neural network for implicit discourse relation recognition[C]//Proceedings of EMNLP2015, 2015:2230-2235.

14. Qin L, Zhang Z, Zhao H. A Stacking Gated Neural Architecture for Implicit Discourse Relation Classification[C]//EMNLP. 2016: 2263-2270.

15. Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

16. Tenney I, Das D, Pavlick E. BERT Rediscovers the Classical NLP Pipeline[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

17. Wang N. T. Research on Improved Text Representation Model Based on BERT[D].Southwest University, 2019.

18. Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proc of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 2001: 282-289.