Named Entity Recognition for Science and Technology Policy Dynamics

Wenjiao Zheng [1], Bolin Hua [1]

[1] *Peking University, No.5 Yiheyuan Road Haidian District, Beijing, 100871, China*

Abstract

Dynamic text of science and technology policy reflects the latest intelligence in the field of science and technology policy. Entity extraction of dynamic text can provide data basis for subsequent downstream tasks such as relationship extraction and knowledge graph construction. This research used RoBERTa+FLAT method to extract the entity, and Transformer structure combining relative position coding and lexical boundary information is used to improve the recognition effect of entity boundary. The experimental results show that the RoBERTa + FLAT compared with the traditional method has a better effect of entity recognition.

Keywords

Named entity recognition, Science and technology policy dynamics, FLAT model, RoBERTa model

1. Introduction

Science and technology policy dynamics refers to media reports on the behavior of policy subjects in the formulation, implementation and supervision of national science and technology policy, including holding meetings, issuing proposals and revising legislation. There are a large number of entities related to science and technology policy in the dynamic text of science and technology policy, including all kinds of organizations, names, policy names, etc. Entity extraction of these entities can provide data basis for further downstream tasks such as entity relationship extraction and knowledge graph construction.

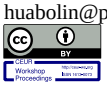
The research methods of NER problem mainly experienced the development from the traditional pattern matching method based on rules, to the method based on statistical machine learning, and then to the deep learning method. Due to the ambiguity of Chinese lexical boundaries, named entity recognition for Chinese domain has received special attention.

The task of named entity recognition in Chinese domain can be divided into two types: character-based model and word-based model. Character-based models have been proved to be better than word-based models, but they cannot utilize lexical information, so there are many studies to improve model performance by integrating lexical information into NER systems. In studies on the fusion of lexical information, Lattice structure has been proved to have great advantages in utilizing word information and avoiding the propagation of segmentation errors [1]. The Lattice matches the characters in a sentence to a dictionary to get the potential words in it, which then results in a grid-like structure. Li et al. proposed a FLAT model of improved Lattice structure [2], which transformed the grid structure into a planar structure composed of several spans, and adopted the fully connected self-attention mechanism to model the long-distance dependence in the sequence.

To sum up, for NER tasks in the Chinese domain, the character-based model using lexical information has higher performance than the word-based model, while the pre-trained language model can dynamically encode input sequences

3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2022), June 20-24, 2022, Cologne, Germany and Online

EMAIL: zhengwenjiao@pku.edu.cn (Wenjiao Zheng); huabolin@pku.edu.cn (Bolin Hua)



and incorporate contextual information into the model. In the selection of data objects for domain entity recognition, based on news, Wikipedia and medical records texts [3], while there are few extraction studies for policy text field. Therefore, the RoBERTa+FLAT scheme is adopted in this paper for dynamic entity extraction of science and technology policies.

2. Method design

The overall research design of the research on dynamic named entity recognition of science and technology policy includes two main modules, namely data collection and pre-processing, model training.

2.1. Data collection and pre-processing

2.1.1. Corpus acquisition

The dynamic text of science and technology policy mainly includes the report of the behavior content related to the policy subject and science and technology policy. The research selected the weekly science and technology policy dynamic reports published on the official website of American Physical Society as the data source, and obtained a total of 3012 reports from 2020-2022 by using the method of web crawlers. In this research, the original English corpus is automatically translated, and the original corpus is converted into Chinese by calling baidu translation interface for subsequent training tasks. As there is no publicly annotated data set in the field of science and technology policy dynamics, four types of entities from the CLUENER2020 Chinese fine-grained named entity recognition data set, including location, company, name and position, are selected for this research.

2.1.2. Dictionary Construction

There is no unified entity specification in the field of science and technology policy at present. Therefore, based on the entity types of OntoNotes fine-grained named entity recognition dataset, nine types of dynamically related entities of science and technology policy are defined, including Government, Company, Research institution, Name, Position, Policy, Conference, Location and Time. Since there is no open domain dictionary in the field of science and technology policy, the research chose to construct a domain dictionary by adding domain words to the general domain lexicon.

2.1.3. Data pre-processing

In this research, the corpus was cleaned and processed by clause processing, and 14,047 sentences were obtained. In this research, 3000 sentences after processing were annotated, and the annotation results were combined with CLUENER annotation dataset to obtain a total of 13, 748 data pieces.

2.2. Model training

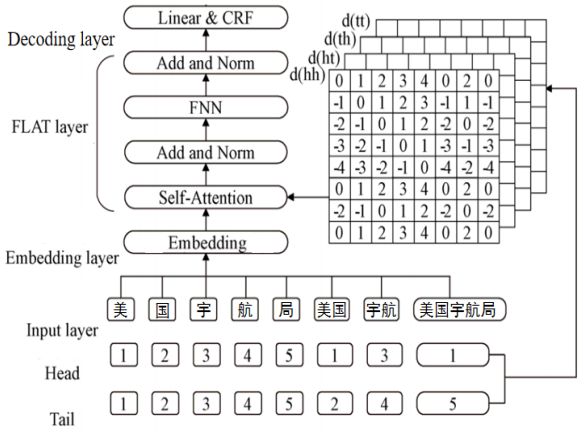


Figure 1: General framework of Bert-Flat model

In the task of named entity recognition, the character-based encoding method is used to avoid the error propagation caused by word segmentation errors, but there are also problems that the semantic information of characters in the vocabulary cannot be used and the lexical boundary cannot be defined. Therefore, the Roberta-Flat model combining character and lexical information is selected for data training. Firstly, the RoBERTa pre-trained language model and Word2vec coding model are used to vectorize the character and vocabulary information respectively, and then the character vector and corresponding word vector are matched and spliced. The spliced vector will be used as the output of the embedding layer to enter the FLAT layer for position coding. The model builds a head position encoding and a tail position encoding for each character and word respectively, and uses Transformer to fully model the encoding results. The output of the FLAT layer will be decoded into

the CRF layer to obtain the predicted results. The specific structure is shown in Figure 1.

3. Analysis and measurement of results

3.1. Analysis of experimental results

In order to verify the effectiveness of relative position coding and the introduction of external word lists, comparative experiments are conducted on BiLSTM+CRF, Iterative expansive convolutional Neural Network (IDCNN) and FLAT, and the test results are shown in table 1.

Table 1
Comparison of experimental results

Training program The F value	Training program The F value
BiLSTM+CRF	78. 48
IDCNN-CRF	70. 52
FLAT	76. 15
RoBERTa+FLAT	78. 99

It can be seen from Table 1 that the RoBERTa+FLAT model achieves the optimal result in the experimental model, which verifies the effectiveness of integrating lexical information and adopting relative position coding. Meanwhile, the addition of RoBERTa model also improves the experimental result to some extent, indicating that the use of pre-trained language model can Simprove the performance of the FLAT model.

3.2. Recognition details analysis

Entity extraction was carried out on 3012 dynamic texts of science and technology policies. After statistical analysis, a total of 21, 320 entities were extracted in the experiment, and the average number of entities extracted from each report was 7. 08.

3.3. System Display

We designed entity label display system based on the entity extraction results. Take reports related to the Infrastructure Investment and Jobs Act as an example. By inputting the entity keyword "Infrastructure Investment and Jobs Act", users can directly retrieve a series of developments related to the entity. Different types of entities are distinguished by different background colors. The specific search result interface is shown in Figure 2.

Figure 2: Display interface of science and technology policy dynamic entity label

4. Conclusion and Discussion

Based on the research of domain named entity recognition, this paper adopts the method of RoBERTa+FLAT integrating lexical information to extract the entity from the dynamic text of science and technology policy. Transformer structure combining relative position coding and lexical boundary information can improve the recognition effect of entity boundary. The experimental results show that the method we used compared with the traditional method has a better effect of entity recognition.

However, the research in this paper also has some limitations. Small-scale annotated datasets affect the training effect of the model. The subsequent research will try to overcome the obstacles of small-scale datasets by using related methods such as domain migration.

Acknowledgments

This work was supported in part by The National Social Science Fundation of China (Number: 17BTQ066).

References

[1] Yue Zhang and Jie Yang, 2018. Chinese NER Using Lattice LSTM. arXiv:1805. 02023. Retrieved from https://arxiv.org/abs/1805. 02023

[2] Xiaonan Li, Hang Yan, Xipeng Qiu and Xuanjing Huang. 2020. Chinese NER using flat-lattice transformer. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 5-10, 2020, Online, 6836-6842. https://aclanthology.org/2020.acl-main.611.

[3] Kainan Jiao, Xin Li, Rongchen Zhu, 2021. A Review of Named Entity Recognition in Chinese Domain Computer Engineering and Applications 57, 16 (May, 2021), 1-15.

[4] Cheng Chen, Cheng Xian Yi and Hua Jin, 2012. Research of Chinese Named Entity Recognition Using GATE Advanced Materials Research 393, 1 (Feb, 2012), 262-264.
DOI: https://doi.org/10.4028/www.scientific.net/AMR.393-395. 262.

[5] Qun liu, Huaping Zhang, Hongkui Yu, Xueqi Cheng, 2004. Chinese lexical analysis using cascaded hidden Markov model Journal of Computer Research and Development 41, 8 (Aug, 2004), 1421-1429.

[6] Zheng Yannan, Tian Dagang, 2018. Text classification based on GloVe and SVM Software Guide 17, 6 (Jun, 2018), 45-48.

[7] Zhuyu Zhang, Feiliang Ren and Jingbo Zhu, 2008. A Comparative Study of Features on CRF-based Chinese Named Entity Recognition. Information Retrieval and Content Security Committee of Chinese Information Society of China, November 15, 2008, Beijing, BJ, 111-117.

[8] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean, 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301. 3781. Retrieved from https://arxiv.org/abs/1301.3781.

[9] Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019. CNN-based Chinese NER with lexicon rethinking. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI '19), August 10-16, 1979, Macao, MO, 4982-4988. https://doi.org/10.24963/ijcai.2019/692.

[10] Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2019. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP), November 3-7, 2019, Hong Kong, HK, 3821-3831. https://aclanthology.org/D19-1396.