Journal of Contemporary Educational Research, 2022, Volume 6, Issue 5
http://ojs.bbwpublisher.com/index.php/JCER
ISSN Online: 2208-8474
ISSN Print: 2208-8466

Overview of Named Entity Recognition

Xing Liu*, Huiqin Chen, Wangui Xia

School of Computer and Information Engineering, Heilongjiang University of Science and Technology, Harbin 150022, China

Corresponding author: Xing Liu, 649325694@qq.com

Abstract: Named entity recognition, as a sub-task of information extraction, has attracted widespread attention from scholars at home and abroad since it was proposed, and a series of studies and discussions have been carried out based on it. This paper discusses the existing named entity recognition technology based on its history of development.

Keywords: Named entity recognition; Information extraction

Online publication: May 30, 2022

1. Introduction

The Sixth Message Understanding Conference (MUC-6) first proposed the named entity recognition [1], which has been widely used in natural language processing tasks. The main task of named entity recognition is to extract specific "proper nouns" from unstructured texts, such as the names of people, places, and institutions, in addition to dates [2]. This is also the noun recognized by the earliest defined tasks of the early MUC-6. As research progresses, scholars have made a more detailed division of the alignment. For names of places, they can be subdivided into countries, provinces, states, and cities [3]. For people, they can also be subdivided into politicians, actors, and other roles [4]. This paper mainly reviews the development history of named entity recognition technology and expounds the progress of named entity recognition research technology, identification methods, and evaluation standards.

2. Technical methods

Named entity recognition refers to entity extraction technology, which is also known as entity extraction [5]. Named entity recognition plays an important role in natural language processing and is the most critical and fundamental step in knowledge extraction. Based on the development history of named entity recognition technology, there are three types of extraction methods.

2.1. Named entity recognition based on rules and templates

The development of early named entity recognition technology is immature, and it is usually performed by constructing rules and templates artificially. In the early 1990s, Rau applied the template- and rule-based method to the named entity recognition task, manually constructed a large number of rules and templates, as well as used heuristic algorithms to successfully identify a company's corporate identity from financial news; the extraction efficiency and accuracy of this method far exceeded manual extraction, in which the evaluation index exceeded 90% [6]. However, this method requires domain experts to customize the rules, consumes time and energy, cannot be applied to other fields, and has poor generalization as well as portability.

2.2. Named entity recognition based on machine learning

After applying machine learning to named entity recognition tasks, domain experts no longer need to manually construct rules or templates, but rather rely on annotated corpus to use these corpora to train models. The representative models include Hidden Markov Model and Conditional Random Field Model. In 2015, Han Chunyan and several other researchers used CRF to extract feature-level, sentence-level, and lexical-level features and input them together with dictionary features into another CRF for entity recognition in the microblog domain [7]. Subsequently, in 2020, Feng Jing and other researchers combined the Hidden Markov Model with lexical features and proprietary rules in the bridge domain to identify bridge entities, with an F1 of 75.7 [8]. The two methods are based on machine learning, so features need to be extracted; hence, there will be error propagation in the model during training. In view of this, scholars gradually began to shift their focus to deep learning.

2.3. Named body recognition based on deep learning

Compared with machine learning, entity recognition based on deep learning is not as high in feature dependence as machine learning, which reduces the dependence on features to a certain extent and solves the error propagation problem in model training. Convolution neural networks (CNN), recurrent neural networks (RNN), and their variant networks are the main application networks of this method.

The application of convolutional neural networks to named entity recognition tasks was originally proposed by Collobert in 2019 [9]. In the same year, domestic scholars added CRF on the basis of CNN and proposed the CNN-CRF model, which was used to extract entities from Chinese electronic cases [10]. Both, accuracy and speed improved after using this model. In 2021, Kong and Zhang added an attention mechanism to deal with the information loss in long sentences based on the fact that the traditional CNN model cannot handle the loss of long-distance information; they proposed a new CNN model, in which the CNN fusion of different convolution kernels and residual structures improved the ability to capture the contextual information of long texts from different dimensions [11].

Besides traditional convolutional neural networks, recurrent neural networks have also been widely used in named entity recognition tasks. Huang and several scholars proposed the application of long-short-term memory network, a variant of the recurrent neural network, to named entity recognition tasks, as well as a series of long-short-term memory network-based models, such as LSTM, BI-LSTM, and others [12]. The features of the text in both past and future directions can be obtained by using the BI-LSTM model. Inspired by Huang, several researchers have carried out a lot of research work on this basis. In another study [13], a bidirectional LSTM model was used to extract 22 entity types, such as diseases, symptoms, body parts, and other components, from electronic case data, and the experimental F1 value reached 80.52%. In 2019, on the basis of BI-LSTM, an attention mechanism was introduced to calculate the importance of key features in text and enhance the extraction ability of text features [14]. Its excellent recognition performance has been proven in named entity recognition on the SIGHAN Bakeoff 3 2016 data set. In 2020, Liu Yupeng and Li Dongdong proposed a new network structure, which is an end-to-end model structure combining CNN and LSTM; both, CNN and CRF were used to obtain word-based representations [15].

In addition to CNN and RNN, in recent years, the application of transformer models to entity extraction has also become a research hotspot for scholars in this field. Transformer is mainly implemented by the attention mechanism. Using transformers for named entity recognition increases the accuracy and shortens the training time. The representatives of using transformers in named entity recognition include the TENER model proposed by Yan [16] and the Transformer-CRF model proposed by Li Bo and other researchers [17]. The former proposes a special transformer structure for named entity recognition tasks, while the latter is used in transformers. On the basis of extracting text features, CRF is introduced for entity classification and recognition. In addition, there is also the ERNIE-BiGRU-CRF model proposed by Zhang Xiao and other

researchers in 2020 [18] and the BERT-BiLSTM-CRF model proposed by Shen Tongping and other scholars in 2022 [19], both of which combine the attention mechanism with RNN; the neural network extracts sentence features and uses the attention mechanism to solve the problem of long-distance dependency, which effectively improves the overall recognition ability of the model. The application of attention mechanism in named entity recognition tasks expands the research direction of named entity recognition.

3. Evaluation criteria

The progress of named entity recognition technology has benefited from numerous evaluations, in which the method used compares the annotation recognition results of linguistic experts with the automatic recognition results by machines [20]. The current mainstream evaluation methods are based on the evaluation committee standards. The MUC is a relatively influential evaluation conference. In the MUC, the performance of the information extraction system is mainly measured based on two evaluation indicators: recall rate and accuracy rate. The recall rate (REC) is equal to the proportion of the results correctly extracted by the system to all possible correct results, whereas the precision rate (PRE) is equal to the proportion of the results correctly extracted by the system to all extracted results. In order to comprehensively evaluate the performance of the system, the weighted geometric mean of the recall rate and accuracy rate, that is the F index, is usually calculated. The formula is as follows:

$$F-\text{measure}=\frac{(\text{beta}^2+1.0)\times PRE\times REC}{(beta^2\times PRE)+REC}$$

where beta is the relative weight of recall and precision. Both are equally important when beta equals to 1; when beta is greater than 1, the accuracy rate is relatively important; otherwise, the recall rate is.

4. Conclusion

Named entity recognition, as a sub-task of information extraction, has received extensive attention from scholars at home and abroad since MUC-6. It has become a research hotspot in the field of information extraction. Beginning from its development history, this paper discusses three types of technical methods: the named entity identification method based on rules and templates, the named entity identification method based on machine learning, and the named entity identification method based on deep learning. At present, the named entity recognition technology for general texts has been established, and many researchers are focusing on the use of deep learning and attention mechanism for named entity recognition.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Grishman R, Sundheim B, 1996, Proceedings of the International Conference on Computational Linguistics, August 5-9, 1996: Message Understanding Conference - 6: A Brief History. Association for Computational Linguistics, Copenhagen, 466-471.

[2] Thielen C, 1995. Proceedings of the EACL-95 SIGDAT Workshop: From Text to Tags, March 27, 1995: An Approach to Proper Name Tagging for German. Dublin, Ireland.

[3] Lee S, Lee G, 2005, Proceedings of the International Joint Conference on Natural Language Processing, October 11-13, 2005: Heuristic Methods for Reducing Errors of Geographic Named Entities Learned by

Bootstrapping. Springer Verlag, Jeju Island, Korea, 658-669.

[4] Fleischman M, Hovy E, 2002, Proceedings of the 19th International Conference on Computational Linguistics, August 24-September 1, 2002: Fine Grained Classification of Named Entities. Association for Computational Linguistics, Taipei, Taiwan, 1-7.

[5] Yu L, Guo Z, Chen G, et al., 2020, Review of Knowledge Extraction Technology for Knowledge Graph Construction. Journal of the University of Information Engineering, 21(2): 9.

[6] Rau LF, 1991, Proceedings of the Seventh IEEE Conference on Artificial Intelligence Application, February 24-28, 1991: Extracting Company Names from Text. IEEE, Miami Beach, FL, USA, 29-32.

[7] Han C, Liu Y, Ju S, et al., 2015, Recognition of Chinese Microblog Names. Journal of Sichuan University (Natural Science Edition), 52(3): 511-516.

[8] Feng J, Li Z, Zhang D, et al., 2020, Bridge Detection Text Named Entity Recognition Based on Hidden Markov Model. Traffic World, 2020(8): 32-33.

[9] Collobert R, Weston J, Bottou L, et al., 2011, Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, 12(1): 2493-2537.

[10] Cao Y, Zhou Y, Shen F, et al., 2019, Research on Named Entity Recognition of Chinese Electronic Medical Records Based on CNN-CRF. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2019(6): 869-875.

[11] Kong J, Zhang L, Jiang M, et al., 2021, Incorporating Multi-Level CNN and Attention Mechanism for Chinese Clinical Named Entity Recognition. Journal of Biomedical Informatics, 116: 103737.

[12] Huang Z, Xu W, Yu K, 2015, Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv, arXiv:1508.01991 (preprint).

[13] Yang H, Li L, Yang R, et al., 2018, Recognition Model of Electronic Medical Record Named Entity Based on Bidirectional LSTM Neural Network. Chinese Tissue Engineering Research, 22(20): 3237-3242.

[14] Ji X, Zhu Y, Li F, et al., 2019, Chinese Named Entity Recognition Based on Attention-BILSTM. Journal of Hunan University of Technology, 33(5): 73-78.

[15] Liu Y, Li D, 2020, Chinese Named Entity Recognition Method Based on BLSTM-CNN-CRF. Journal of Harbin University of Science and Technology, 25(1): 115-120. DOI: 10.15938/j.jhust.2020.01.017

[16] Yan H, Deng B, Li X, et al., 2019, TENER: Adapting Transformer Encoder for Named Entity Recognition arXiv, arXiv:1911.04474 (preprint).

[17] Li B, Kang X, Zhang H, et al., 2020, Named Entity Recognition of Chinese Electronic Medical Records Using Transformer-CRF. Computer Engineering and Applications, 56(5): 153-159.

[18] Zhang X, Li Y, Wang D, et al., 2020, Named Entity Recognition Based on ERNIE. Intelligent Computer and Application, 10(03): 21-26.

[19] Shen T, Yu L, Jin L, et al., 2022, Research on Chinese Entity Recognition Based on BERT-BILSTM-CRF Model. Journal of Qiqihar University (Natural Science Edition), 38(01): 26-32.

[20] Sun Z, Wang H, 2010, Overview on the Advance of the Research on Named Entity Recognition. New Technology of Library and Information Service, 26(6): 42-47.