Overview of Naming Entities Based on Natural Language Processing

Tao Li[1], Xuhan Jin[2]

[1]School of Innovation and Entrepreneurship, Huaiyin Institute of Technology, Jiangsu, China
[2]School of Mechanical and Electrical Engineering, Anhui Jianzhu University, Anhui, China

Abstract: *Named entity recognition is an important research direction of natural language processing. This paper first reviews the development process and main stages of named entity recognition, then expounds the research contents and methods of named entity recognition, and puts forward the key and difficult points of named entity recognition, The development process of named entity recognition is from the initial rule-based and dictionary based method to the later statistical learning method, and then to the mixed learning method and some popular learning methods. Finally, the development of named entity recognition is prospected.*

Keywords: *Name entity identifying, Natural language processing, Artificial intelligence.*

1. Introduction

Name entity identification is a research hotspot in natural language processing. In the early days of the development of artificial intelligence, people put forward the ultimate goal of artificial intelligence to achieve human intelligence level [1], in order to achieve this goal, it is necessary to have a full range of databases for natural language processing. However, due to the complexity and professionalism of knowledge, the construction of the database has a long way to go. The entity in the text is usually the key to understanding the text, which contains rich information. Name entity identification is to find name entities in the text and classify them, such as place names, name, time and other entities. The development of naming entity identification technology is from the initial rule-based approach to statistical methods, and then to the combination of the two, the last is some of the hot research methods. Name entity identification plays an important role in the field of natural language processing, such as machine translation [2], intelligent question and answer [3], build database [4-5], etc.

The name entity recognition was first proposed by Rau et al [6] scholars. They first combined manually prepared rules and inspiration ideas, which could automatically identify naming entities of the company name class from text. This method was based on rules, and there was still a certain drawback, which was limited to the company's name entity, not suitabled for other naming entities categories.

In the early stage of research, naming entity identification was mainly for the study of English entity identification. Due to the words of the English itself are separated by spaces, it was not necessary to distinguish sentences, bringing great convenience to research. For Chinese studies, due to the semantic and rich characteristics of Chinese, the Chinese text must be analyzed before the naming entity recognition, which led to complicated research on Chinese. In the early studies of Chinese, Name entity identification was mainly to conduct research on people's name and place name, such as Sun Massan [7] scholars calculated the probability of using statistics for human name and place name. Zhang Xiaoheng et al [8] scholars analyzed the names of universities through artificial rules.

In the early named entity recognition research, the research of English texts mainly: Bikel et al. [9] scholars put forward a research method for English text according to the hidden Markov model. Liao et al [10] scholars submitted research by semi-supervised learning according to the conditions of the airport model. RatinoV et al. [11] scholars put forward a method of text training model, which is highly efficient.

2. Research Content and Difficulties of Naming Entity Identification

2.1. Research Content of Name Entity Identification

The research subject identified by nomenclastries mainly includes 3 large classes (physical classes, time classes, digital classes) and 7 categories (name, place name, institution name, time, date, currency, percentage). But in actual research, since naming entities need to be determined in specific applications. The naming entity of the time class and digital classes typically can get a good effect by manually matching, and the physical classes are more complicated, so most of the research is directed to this. From the development of research, various named entities are identified by a separate study of a single study to a unified approach. Such a method can take into account the common characteristics of the entity nomenclature, to some extent, solve the problem of ambiguity, but cannot completely distinguish different entities, and the identification of the text needs to be improved.

There is a weak correlation in the field of nomencutical entities, and there are similar features in different areas, but in the process of transitioning between fields, there is often a bad effect, mainly because of different fields Name entities have different grammatical features.

2.2. Difficulties of Naming Entity Identification

There are many scholars who believe that the research on naming entities is very fully, and there is a high accuracy of many named entities, there is no need to continue research, but in actual research, naming entities are still facing challenges:

Name entity recognition is currently good in specific sectors, such as some people names, place names and institutions in the news, but these technologies cannot be applied to other fields such as medical, military, and biological. The diversity of natural language has brought great challenges to the development of nomenclature, in different technologies, cultural contexts, the extension of naming entities is different, which is a problem that the naming entity identification technology needs to be solved. The naming entity has an expression unclear, referring to an unknown phenomenon, which requires the need to understand the meaning of the context, and more accurately perform nomencutive entities more accurately by understanding the meaning of naming entities in context.

3. Research Methods for Named Entity Identification

Name entity identifies an important branch of natural language processing. The object of the study is a nomenclature that identifies a naming entity in the text and classify them. The development process of naming entity identification technology is from the initial rules and dictionary methods, to traditional statistical learning, reappearing to the mixing method, and then to the current figure neural network, semi-supervisor learning and other popular methods.

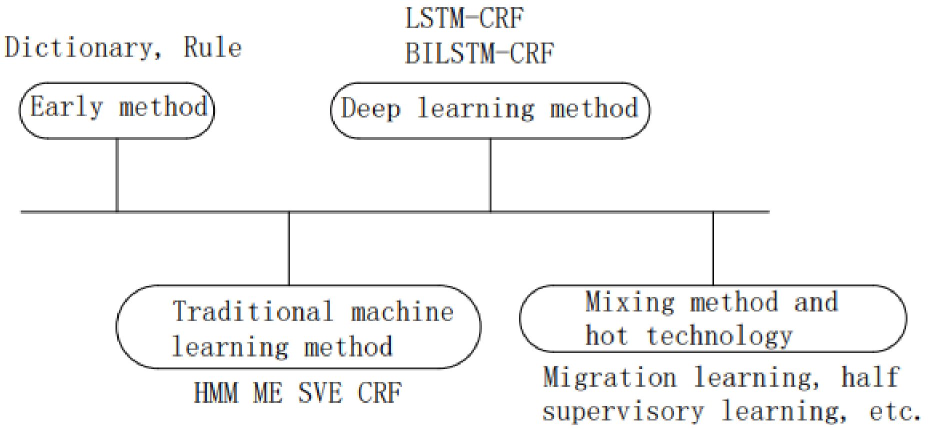
Dictionary, Rule
LSTM-CRF
BILSTM-CRF
Early method
Deep learning method
Traditional machine learning method
Mixing method and hot technology
HMM ME SVE CRF
Migration learning, half supervisory learning, etc.

Figure 1: Development trend of naming entity identification technology

3.1. Methods Based on Rules and Dictionaries

Based on rules and dictionary is the first research method for naming entities identify, mainly by artificial way, based on the rule template built by the database or a special dictionary, then use the matching method to make naming entity identification.

The name entity recognition was first proposed by the Rau et al [12] scholar, and they first combined artificially prepared rules and inspiration ideas, which can automatically identify naming entities of the company name class from text. This method is based on rules, and there is still a certain drawback, which is limited to the company's name entity, not suitable for other naming entities categories.

3.2. Based on Statistics

Named entity identification mainly includes the following models: Hidden Markov Model, HMM, Maximum Entropy [13], Support Vector Machine, and Conditions Compensation [14].

Among these models, the maximum entropy model is relatively close, compared to other models are very versatile, but this model has certain disadvantages, and the complexity of training is high. Since there is a clear normalization, consumption Human and time will also increase. The hidden Markov model is a probability that the text is directly modeling, the words in the statistical text are co-developed in a certain frequency. Support vector machines and maximum entropy models are improved with respect to the hidden Markov model, but in terms of speed, the hidden Markov model has an advantage. The condition of the airport model also has a slower, time long, but the condition random is still the mainstream model of traditional machine learning, because in the process of sequence labeling, the conditional random model can use the text structure information and Combined with the context information, make the results more accurate.

There are also scholars to improve the traditional machine learning methods by adjusting model adjustment, making the accuracy and recall rate of the model. Culotta and McCallum [15] calculate the confidence score of the phrase extracted from the CRF model, and these scores are used to sort and filter entity identification. Carpenter [16] Calculate the condition probability of the phrase level from the HMM and attempt to increase the recall of the naming entity by reducing the thresholds of these probabilities. For a CRF model of a given training, MINKOV, etc. [17] scholars judge whether it is a naming entity through the weight of the fine-tuning feature, changing the weight may reward or punish the entity identification during the CRF decoding process.

3.3. Method and Mixing Method Based on Deep Learning

With the continuous development of deep learning, the research focus of naming entities has turned to deep learning, which does not require knowledge of characteristics. Collobert et al.[18]scholars first propose a nomenchatic entity identification method based on a neural network. Each word in this method is fixed, but the limit is limited to the valid information of the long distance of the word. In order to solve this problem, Chiu and Nichols [19] provide a two-way LSTM-CNNS structure, which can automatically detect the characteristics of words. MA and Hovy [20] further extend to the BILSTM-CNNS-CRF structure, which adds a CRF module to optimize the output. LiU et al [21] proposes a neural language model of LM-LSTM-CRF, incorporating multiple language models into a common framework to quantify the character. These models have functions that automatically learn from the data, which can recognize new nomenchat entities well.

As the research continues, naming entity identification techniques are more inclined to deal with the text using a mixing method, and is usually used to perform a pre-processing for text based on rules, and then processed by statistics. The main categories of the mixing method are as follows:

The first: the superposition of a single model is used to achieve the effect of mixing, such as superposition of the hidden Markov model, and then treated Chinese using the superimposed method.

Second: Binding based on rules and statistical methods, adding rule-based artificial methods or joining statistical methods in a statistical approach, combining manual and machines, achieving mixing effects.

The third type: Mix the various models and methods, this method should take into account how to effectively combine the front and rear methods, and efficiently apply the processing data of the previous model to the next data.

Lin et al [22] scholars have enhanced the accuracy and recall of the model based on rules and statistical and statistical combinations. Greenberg et al. [23] scholars proposed a single CRF model that uses the heterogeneous signature to name entity identification, this method has practicality to the domain data set of balance label distribution. AuGenstein et al [24] uses label to quantify the further information between tasks. Beryozkin et al. [25] It is recommended to use a given label hierarchy to learn a neural network to share their label layers in all labels. It has achieved very excellent performance. In recent years, hot research technology such as neural network, migration learning, far-reaching study, etc., is also the current mainstream research direction.

4. Conclusion

Name entities identify applications in other disciplines are also an important research direction. Effectively apply existing methods on a variety of fields to help a variety of disciplines get the name entity they have focused, which is the meaning and value of naming entity identification research.

References

[1] Chinchor N,Robinso N P.MUC-7 Named Entity Task Definition [C] //Proceedings of the 7th Conference on Message Understanding,1997,29: 1-21.

[2] Babych B, Hartley A. Improving Machine Translation Quality with Automatic Named Entity Recognition [C] //Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT Through Other Language Technology Tools, Resource and Tools for Building MT at EACL 2003, 2003.

[3] Bordes A, Usuniern, Chopra S, et al. Large-scale Simple Question Answering with Memory Networks [J]. arXiv preprint arXiv: 1506.02075, 2015.

[4] Riedel S,Yao L, Mccallum A, et al. Relation Extraction with Matrix Factorization and Universal Schemas [C] //Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013: 74-84.

[5] Shen W, Wang J, Luo P, et al. Linden: Linking Named Entities with Knowledge Base Via Semantic Knowledge [C] //Proceedings of the 21st International Conference on World Wide Web, 2012: 449-458.

[6] Rau L F. Extractingcompany name from text [C] //Proceedings of the sevesth IEEE Conference on Artificial Intelligence Application. IEEE, 1991, 1: 29-32.

[7] Sun Maosong, Huang Changning, Gao Haiyan, etc. Automatic Identification of Chinese Names [J]. Chinese letter Journal, 1995, 9 (2): 16-27.

[8] Zhang Xiaoheng, Wang Lingling. Identification and Analysis of Chinese Institution Name [J]. Chinese information Report, 1997, 11 (4): 21-32.

[9] Bikel D M, Schwarta R, Weischedel R M. An Algorithm That Learns What's in a name [J]. Machine Learning Journalland Learning, 1999, 34 (1-3): 211-231.

[10] Lia W, Veeramachaneni S. A Simple Semi - Supervised Algorithm for named Entity Recognition [C]. In: Proceedings of the NaaCl HLT 2009 Workshop on Semi - Supervised Learning for Natural LANGUAGE Processing. 2009: 58-65.

[11] Ratinov L, Roth D. Design Challenges and Misconceptions in Named Entity Recognition [C]. In: Proceedings of the 13th Confer Ence on Computational Natural Language Learning. 2009: 147-155.

[12] Xie R, Liu Z, Jia J, et al. Representation Learning of Knowledge Graphs with Entity Descriptions [C] // Thirtieth AAAI Conference on Artificial 258 Radio Communications Technology Vol.46 No.3 2020 Intelligence, 2016.

[13] Ratnaparkhi A. A Maximum Entropy Model for Part-Of - Speech Tagning [C] // Conference on Empirical Methods in Natural Language Processing, 1996: 133-142.

[14] Lafferty J, McCallum A, Pereira F C N. Condi Tional Random Fields: probabilistic models for segmenting And labeling sequence data [c] //proceedings of the 18th International Conference On Machine Learning 2001 (ICML 2001): 282-289.

[15] Culotta A, McCallum A. Confidence Estimation for Information Extraction [C] // Proceedings of HLT-NaaCl 2004: SHORT PAPERS, 2004: 109-112.

[16] Carpenter B. Ling pipe for 99. 99% Recall of Gene Mentions [C] // Proceedings of the second biocreative CHALLENGE EVALE WORKSHOP. Biocreative, 2007, 23: 307-309.

[17] Minkov E, Wang R C, Tomasic A, et al. Ner Systems That Suit User's Preferences: Adjusting THE Recall-precision track - off for entity extraction [c] //Proceedings of the human language technology Conference of the NaaCl, Companion Volume: Short Pers. 2006: 93-96.

[18] Collobert R, Weston J, Bottou L, et al. Natural Language processing (almost) from scratch [j]. Journal of Machine Learning Research, 2011, 12 (AUG): 2493-2537.

[19] Chiu J P C, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNS [J]. TRANSACTIONS OF THE Associ at Computational Linguistics, 2016, 4: 357-370.

[20] Ma X, Hovy E. End - to - End Sequence Labeling Via Bi-Directional LSTM-CNNS-CRF [J]. Arxiv Preprint Arxiv: 1603.01354, 2016.

[21] Liu L, Shang J, Ren X, et al. Empower Sequern Labe Ling with Task - Aware Neural Language Model [C] //Thirty-Second Aaai Conference on Artificial Intelligence. 2018.

[22] Lin Y, Tsai T, Chou W, et al. A Maximum Entropy Approach to Biomedical named Entity Recognition [C]. In: Proceedings of The 4h ACM Sigkdd Workshop on Data Mining in BioInformatics. 2004.
[23] Greenberg N, Bansal T, Verga P, et al. Marginal Likelihood Training of Bilstm-CRF for Biomedical Named Entity Recognition from Disjoint Label Sets [C] // Pro Ceedings of the 2018 Conference ON Empirical Methods in Natural Language Processing, 2018: 2824-2829.
[24] Augenstein I, Ruder S, Sogaard A. Multi – Task Learning of Pairwise Sequence Classification Tasks over Disparate Label Spaces [J]. Arxiv Preprint Arxiv: 1802.09913, 2018.
[25] Bryozkin G, Drori Y, Gilon O, et al. A joint Named - Entity Recognizer for Heterogeneous Tag – Sets Using a tag hierarchy [j]. Arxiv Preprint Arxiv: 1905.09135, 2019.