Information Extraction Research Review

Dengyun Zhu, Qi Guo, Dongjiao Zhang and Fucheng Wan*

Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest MinZu University, Lanzhou, China

*Corresponding author's email: loda0622@163.com

Abstract. Information extraction is to extract the required specific entity, event, relationship and other information from a large number of texts and store it in a structured form, which occupies an important position in the field of natural language. Information extraction helps people get the information they need efficiently, avoids a lot of complicated work, and saves manpower and material resources. In this paper, named entity recognition and so on are described and summarized in detail, as well as the prospect of information extraction.

Keywords. Named entity recognition, Relation extraction, Event extraction

1. Overview of information extraction
In the era of information explosion, people who want to find the information they need from a lot of data need to use information extraction. The early development of information extraction was slow until the TIPSTER text plan proposed by the US government in the 1980s, which accelerated the development of information extraction technology. The organization and convening of MUC played a decisive role in promoting the development of information extraction technology and determined the evaluation tasks and standards. Since 1999, automatic content extraction (ACE) evaluation conference has been organized to further promote research. Compared with the MUC, it supports three types of text input, no longer limiting fields and scenarios, but customizing a set of more systematic information extraction framework from the semantic perspective. Subsequent academic conferences such as multilingual entity evaluation task (MET) and summarization analysis conference (SUMAC) have also promoted the development of information extraction technology. Information extraction is a structured processing of information extracted from natural language texts [1]. For example, extract consumer information from web pages, extract relevant financial information from a large number of financial statements, etc. The extraction process of natural text information is generally shown in Figure 1.
2. Named entity recognition

Named entity recognition is to identify and classify phrases [2] such as time, place, name and proper noun in the text to be processed, which is mainly used in information extraction, machine translation, question answering system and so on. Jing Sifeng et al. [3] proposed a rule-based extraction method for chemical accident cases, which can accurately extract the key information in chemical accident cases for storage and summary, so that people can effectively use their case information. In order to enable job seekers to understand the employment situation and quickly obtain useful employment information, Zhang Xiaoluan et al. [4] proposed Web recruitment information extraction and optimized the extraction results based on semantic extraction. Through constructing the cascaded conditional random field model, Liang Lirong et al. [5] conducted refined text processing for various electronic medical record information of respiratory diseases.

2.1. Methods for named entity recognition

Named entity recognition technology can be divided into rule-based and dictionary-based method, statistics-based method, mixed method, etc. The rule-based method [6] mostly uses rules written by hand and has dependencies on the knowledge base and dictionary created. The method is time-consuming and error-prone, the system portability is not good and the rules in different domains need to be rewritten. Methods based on statistical machine learning [7] mainly include: hidden markov model, maximum entropy, support vector machine, conditional random field, etc., as shown in Table 1. Statistics-based method also relies on corpus, while there are few corpus that can be applied in different fields on a large scale. The hybrid method [8] is to combine the method of rules and dictionary with the method based on statistics. The core of this method is to introduce some rules into the machine learning method and filter the knowledge according to the rules, which solves the problem of large state search space.

Compared with the traditional rule-based and statistical methods that rely on the design of feature templates, the deep learning method can automatically obtain the feature information of serial data through the neural network model, and describe the problem more accurately. CNN (convolutional neural network) and RNN (recurrent neural network) are two typical neural network models. On this basis, researchers have also constructed LSTM, RCNN, C-LSTM, BILSTM and other structures. This method is widely used in the field of natural language. Wang Ziniu et al. [9] proposed a classification model (LAC) based on semantic enhancement and feature fusion. The experiment shows that this model is more effective than the separate CNN model and the LSTM-CNN model that firstly extracts semantic information with LSTM and then perform local feature extraction. Li Ke et al. [10] applied long short-term memory (LSTM) to skip-thought vectors model (STV), introduced a deeply hierarchical bi-directional feature-extraction model by combining with bi-directional information mining method, user preference mining method and deeply hierarchical model, and combined it with singular value decomposition (SVD) to realize commodity recommendation. Wang Dong et al. [11] proposed a Chinese named entity recognition model, by using bidirectional quasi-recurrent neural network (BQRNN) and conditional random field (CRF), sequence context information can be annotated and key information output accuracy can be improved.
Table 1. Models based on statistical learning methods

| Model                      | Principle                                                                 | Benefits                                                                                                    | Drawbacks                                                                                                           | Application                                                    |
|----------------------------|---------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------|
| Hidden Markov Model        | The model state transition process is unknown and observable events are random functions of the state | It has high efficiency and real time to solve the named entity class sequence                                 | The objective function and the prediction objective function do not match                                          | Voice recognition, sequence analysis, quantitative trading, face recognition, etc |
| Maximum Entropy            | When the unknown part of the distribution is known, the event distribution is reflected by the maximum entropy (the most uncertain random variable) | High accuracy, flexible feature selection and strong portability                                           | The iterative process requires a large amount of computation, has sparse data, and is more dependent on corpus | The part of speech tagging, phrase recognition, anaphora resolution, etc |
| Support Vector Machine     | The decision plane with the maximum geometric spacing is found in the training set | Simplified classification regression problem with good "robustness"                                        | Large-scale training samples are difficult to implement                                                          | Text classification, image recognition, speech signal processing, time series, etc |
| Conditional Random Field   | A conditional probability distribution model for annotating and analyzing sequence structures | A globally optimal annotation framework with flexible features                                             | Slow convergence speed and long training time                                                                  | Lexical analysis of Chinese word segmentation and part-of-speech tagging |

3. Relationship extraction

On the basis of entity recognition, relationship extraction is generally a binary classification problem to determine whether there is a relationship between entities in a sentence. Two entity types and the semantic relationship between entities form a triplet <entity1, relation, entity2>, and the relation extraction algorithm is generally output in the form of triplet. Chen Yanguang et al. [12] converted the unstructured criminal judgment documents with natural language records into structured triplet form to facilitate the storage and query of legal information. The relationship extraction methods can be divided into supervised relationship extraction, semi-supervised relationship extraction, remote supervised relationship extraction and unsupervised relationship extraction.

3.1. Supervised learning method

Supervised learning takes it as a classification problem, trains classifiers based on existing data, and uses trained classifiers to predict relationships [13]. When we are answering a question, we can use supervised learning method if we can manually mark some samples with naked eyes, even if there is no training sample. The method of feature extraction is to generate lexical, syntactic and semantic features from the text and train different classifiers in the form of feature vectors. The method based on kernel function obtains the similarity between relation instances by calculating the kernel function.
between objects and carries on the structural representation. The method relies on manual labeling corpus, which is time-consuming and labor-intensive.

3.2. Semi-supervised learning method
The semi-supervised learning method [14] obtained the classification model through iterative training of a small amount of labeled corpus and a large number of unlabeled samples, which did not require a large amount of manual labeled corpus. For example, bootstrapping method, active learning method [15], and tag propagation method [16], etc. The general extraction process is shown in Figure 2.

![Figure 2. General flow of semi-supervised relationship extraction method](image)

3.3. Unsupervised learning method
Unsupervised learning method: As both supervised and semi-supervised entity relationship extraction rely on manual tagging corpus, to solve this problem, an idea based on clustering was proposed [17]. Instead of relying on relational tagging corpus between entities, entities with high similarity were clustered according to context. For example, we know nothing about paintings, but after seeing many works, we can at least know the category of some works.

3.4. Remote supervised relationship
Mintz et al. proposed the remote supervision method, which reduced the cost of manual labeling and could automatically extract a large number of entity pairs, thus expanding the scale of knowledge base [18]. However, based on the basic assumption of remote monitoring, the relationship between entity pairs of massive data will be mislabeled, resulting in noise data. For this problem, methods such as introducing multi-example learning [19], combining the Attention mechanism [20], fitting noise and modeling as reinforcement learning problem [21] are proposed. Comparison of 4 relation extraction methods is shown in Table 2.

| Extraction method               | Implementation | Artificial dependence | System portability | Improved method                                      |
|--------------------------------|----------------|-----------------------|--------------------|------------------------------------------------------|
| Supervised method              | classification | strong                | weak               | Improved rules, features, and kernel functions       |
| Semi-supervised method         | classification | moderate              | moderate           | Improved pattern expansion and noise filtering algorithm |
| Unsupervised method            | clustering     | weak                  | strong             | Extend the features and improve the clustering algorithm |

Table 2. Comparison of relation extraction methods
4. Event extraction

In information extraction, an event refers to an objective fact composed of specific people, objects and actions at the specific time and place. The elements of an event include the event trigger word, event type, event argument and argument role. Event extraction [22] identifies specific types of events from unstructured information and determines and extracts relevant information, which is widely used in information search, intelligent question-answering, machine abstracts and other fields. The event extraction framework is shown in Figure 3. Event extraction is divided into meta-event extraction and subject event extraction.

![Figure 3. Event Extraction Framework](image)

4.1. Meta-event extraction

Meta-event extraction represents the occurrence or state change of an action, including the main components, time, place, character and so on, that participate in the action. Meta-event extraction can be divided into two ways: meta-event extraction based on pattern matching and machine learning. The former refers to event detection and information extraction under certain context constraints. Li Zhangchao et al. [23] built trigger word lists based on pattern matching, filtered to obtain candidate war sentence sets, and built the basic framework system of war events in Zuozhuan based on the obtained structured data. The latter transforms event detection and element recognition into a classification problem through the selection of classifiers and eigenvalues. Luo Xin [24] proposed an attention-based neural network event extraction system that enables users to quickly obtain effective information from a large amount of information. Yu Hui et al. [25] proposed to extract drug treatment events based on BiLSTM-CRF Chinese clinical guidelines event extraction method, which can significantly improve the automatic treatment efficiency of clinical guidelines.

4.2. Subject event extraction

Subject events are extracted around the central topic to get the events related to the subject event. Subject event extraction can be divided into: subject event extraction based on event framework and ontology. The former constructs a structured event framework to extract and summarize the subject event information. Yang Bo et al. [26] constructed a hierarchical toponym address recognition model based on emergency information matching. Wang Yinghuan [27] completed the information extraction of the case subject in the judicial judgment by designing a structured representation framework to describe the case subject. The goal of the latter is to give the relationship between terms from different levels of formal patterns by acquiring and understanding ontology-related knowledge. Ontologies are a great tool for describing subject events. Zhang Zixuan [28] based on the event ontology library, found out the suitable scene and model, and generated the animation content with rich plots by analyzing the theme and template. The ontology-based web page data extraction technology proposed by Guo Xiaoyu [29] receives information from web page resources and establishes a connection between the extracted information results and ontology to enrich ontology resources and improve web page adaptive ability.
5. Acknowledgments
Information extraction (IE) is the extraction of useful information from raw data, such as entities, relationships, events and attributes. Information extraction technology has been widely used, such as classical rules and statistical methods, sequential labeling methods, neural network models and so on. Information extraction is still a great challenge and has not been well solved. Domain named entity recognition has limitations. Data from different domains often have unique characteristics of the domain, so model training is difficult to be carried out. Due to the different granularity of knowledge representation and the lack of normative constraints, there are ambiguities in the representation of named entities. In addition, due to the inability to obtain large-scale marker corpora, the remote supervised method is proposed, which causes problems of mislabeling and error propagation, and the model performance is difficult to evaluate. How to mitigate mislabeling and error propagation is also an important subject for information extraction.

6. References
[1] Information Technology - Data Acquisition 2020 Studies from Masaryk University Further Understanding of Data Acquisition (Traffic and log data captured during a cyber defense exercise). Information Technology Newsweekly
[2] Information Technology - Data Extraction 2020 Recent Findings from Josef Stefan International Postgraduate School Has Provided New Information about Data Extraction (A Survey of Named-entity Recognition Methods for Food Information Extraction) Computer Technology Journal
[3] Jing S F, Xiong G, Liu X W, Gong X Y, Hu B 2019 Research on key information extraction of chemical accident cases Industrial safety and environmental protection 45(08) pp 61-65
[4] Zhang X L, Wang X F 2019 Research on key technologies of semantic Web recruitment information extraction Application of micro computer 35(06) pp 69-70+77
[5] Liang L R, Li C W, Shen Y, Zhou L J, Jing X, Tong C H 2019 Text information extraction of electronic medical record based on the layered conditional random airport model Computer application and software 36(10) pp 47-54+112
[6] Huo J H 2020 Research review on multi-label text emotion classification Modern computer science (14) pp 27-30
[7] Li Y X 2019 Research on statistical machine learning method for false user identification in mobile APP Heilongjiang University
[8] Yang B 2018 Design and implementation of vehicle-borne voice control system based on oral understanding of rules and statistical methods Hebei University of Science and Technology
[9] Wang Z N, Wang H J, Gao J L 2020 Text classification based on semantic enhancement and feature fusion Software 41(01) pp 211-215
[10] Li K, Chen G P 2020 Amazon product recommendation based on deep semantic features of text Computer science 47(02) pp 65-71
[11] Wang D, Li Y G, Zhang X, Pu X Z 2020 Chinese named entity recognition based on quasi-cyclic neural networks Computer engineering and design 41(07) pp 2038-2043
[12] Chen Y G, Wang L, Sun Y Y, Wang Z Z, Zhang S C 2020 Legal text oriented triple extraction technology research. Computer engineering pp 1-9 https://doi.org/10.19678/j.issn.1000-3428.0057677
[13] Cai J J 2019 Progress of text information extraction in AI Journal of hubei university of engineering 39(06) pp 65-72
[14] Wang C D, Xu J, Zhang Y 2020 Summary of entity relationship extraction. Computer engineering and applications 56(12) pp 25-36
[15] Liu J W, Xu J, Zhang L P, Rui W K 2017 Social relation extraction of characters based on label communication and active learning Computer engineering 43(02) pp 234-240
[16] Li Z Z, Feng D W, Li D S, Lu X C 2020 Learning to select pseudo tags: A Semi-supervised learning method for named entity identification Frontiers of Information Technology & AMP; Electronic Engineering 21(6) pp 903-917
[17] Zhang X W, Lv M Q, Li H 2020 Cross-domain pedestrian recognition based on local semantic feature invariance Journal of Beijing university of aeronautics and astronautics pp 1-10 https://doi.org/10.13700/j.bh.1001-5965.2020.0072
[18] Bai L, Jin X L, Xi P B, Cheng X Q 2019 Research review of relationship extraction based on remote supervision Chinese journal of information technology 33(10) pp 10-17
[19] Xu J G, Xiao H F, Zhao H 2020 Text classification algorithm based on multi-example learning framework Computer engineering and design 41(04) pp 1017-1023
[20] Chen Y H, Wang Z 2020 Remote monitoring relationship extraction based on attention mechanism and residual network Computer and digital engineering 48(04) pp 909-913
[21] Li Z P 2019 Research on manufacturing process modeling method based on multi-agent reinforcement learning Qilu University of Technology
[22] Duc-Thuan V, Feras A, Ebrahim B 2020 Extracting temporal and causal relations based on event networks Information Processing and Management 57(6)
[23] Li Z C, Li Z K, He L 2020 Research on extraction technology of war incidents in Zuozhuan Books and intelligence work 64(07) pp 20-29
[24] Luo X 2020 Design and implementation of neural network event extraction system based on attention mechanism University of Electronic Science and Technology of China
[25] Yu H, Xu C, Liu Y R, Fu Y W, Gao D P 2020 Extraction of treatment events based on BiLSTM-CRF clinical guidelines in Chinese Chinese journal of medical books and information 29(02) pp 9-14
[26] Yang B, Wang J Z, Ma W J, Mao X 2017 Emergency geographic information extraction for event framework Science of surveying and mapping 42(12) pp 83-87
[27] Wang Y H 2019 A study on thematic event extraction of the case facing judicial judgment Hefei University of Technology
[28] Zhang Z X 2017 Mobile phone 3D animation event-based plot planning Computer system application 26(05) pp 1-8
[29] Guo X Y, Peng H 2020 Discussion on ontology-based web page data extraction technology Chinese Journal of Multimedia and Network Teaching (the first issue) (07) pp 4-5