

Researches Advanced in the Development and Application of Information Extraction

Yijiao Liu*

School of Fujian Normal University, Fuzhou, China

*Corresponding author: 15001020343@xs.hnrit.edu.cn

Abstract. Natural language processing (NLP) is an interdisciplinary subject of linguistics and computer science, which processes the language used for human communication into a machine language that can be understood by machines. The task of information extraction is to obtain target information accurately and quickly from a large amount of data and improve the utilization of information. With the rapid development of Internet applications, how to quickly and accurately analyze the really useful information in these text data is particularly critical and urgent. Information extraction has become an important branch of natural language processing. Thanks to the rapid development of deep learning, the performance of information extraction has made breakthroughs in recent years. In this paper, based on the detailed literature analysis, this paper first review the development of information extraction. Secondly, the research progress of key technologies of information extraction is summarized from four aspects: named entity recognition, anaphora resolution, relationship extraction and event extraction. Finally, we analyze some main problems of information extraction and predict the research trend of information extraction.

Keywords: Natural language processing; Information extraction; Relation extraction; Lexical analysis.

1. Introduction

With the rapid development of Internet applications, the data (images, text, video, voice, etc.) that can be obtained through the network is also growing exponentially. In these massive data, text, as the main form of information bearing, contains a lot of profound and valuable knowledge. How to quickly and accurately analyze the really useful information in these text data is particularly critical and urgent. In this context, natural language processing technology has developed rapidly, and its basic tasks can be roughly divided into natural language understanding and natural language generation. The research field of language generation focuses on making machines speak like humans, while the purpose of natural language understanding is to make machines understand people. Extracting key information from text data quickly and accurately provides a key modeling basis for the realization of the above two tasks. As a key issue in natural language processing research, information extraction (IE) has attracted a large number of scholars' research interests.

Information extraction aims to extract structured information from unstructured text, such as symptoms, test results and a series of other information from patients' medical records. These information are extracted and stored in the database in a structured way, which is convenient for users' subsequent query and use. In the early days, most information extraction systems used rule-based methods, which rely on manual formulation of extraction templates. These classical information extraction methods can be predicted and explained, while their performance are usually limited due to poor portability or difficulties to summarize effective rules in many scenes. Since the 1990s, statistical models have become the mainstream method, which usually formalize the information extraction task as the prediction from text input to specific target structure by using statistical models to model the mapping between input and output, and using machine learning methods to learn model parameters. In the process of the era of deep learning, relevant researchers began to study how to make good use of one of the features of deep neural networks, which is called automatic learning feature, so as to avoid the problem of error accumulation when extracting features using traditional natural language processing. With the deepening of research, the paradigm of large-scale pre training

language model and downstream task fine-tuning has become the mainstream, and the performance of information extraction model based on deep neural network has been greatly improved. At present, the main tasks of information extraction include entity extraction, anaphora resolution, relationship extraction and event extraction.

(1) Entity extraction, also known as named entity recognition (NER), refers to extracting entities with specific meanings from text, such as person names, place names, organization names, proper nouns, etc. which plays an important role in the question answering system, translation system and knowledge graph. The early ner method takes the matching of pattern and string as the main means, and relies on linguists to manually construct rule templates and select specific features (including statistical information, punctuation, deixis, direction words, head words, etc.). This kind of method requires a lot of manpower to build a language model with a long system cycle, slow knowledge update and poor portability. With the application of machine learning, statistical methods are proposed, mainly including hidden Markov model (HMM), Maximum Entropy Markov model (MEMM), support vector machine (SVM), conditional random field (CRF). Statistical methods have higher requirements for feature selection and greater dependence on corpus. Compared with machine learning feature engineering, deep learning representation learning has great advantages in feature learning. Sentences are embedded into cnn-crf to automatically learn features, classify entities, and the extracted lstm-crf and bilstm-crf models improve entity recognition to a new level.

(2) Anaphora resolution. Anaphora resolution refers to the problem of determining which noun phrase the pronoun points to in the text. Anaphora resolution mainly includes three parts: dominant pronoun resolution, zero pronoun resolution and co referential resolution. Explicit pronoun resolution refers to the fact that the current anaphora has a close semantic correlation with the words, phrases or sentences (sentence groups) appearing in the context. Anaphora depends on the context semantics, and may refer to different entities in different language environments, with asymmetry and non transitivity. Zero pronoun resolution is the process of restoring zero pronoun to refer to the antecedent linguistic unit, sometimes referred to as ellipsis recovery. Co-reference resolution mainly refers to two nouns (including pronouns and noun phrases) pointing to the same reference in the real world. This kind of reference is still valid out of context.

(3) Relationship extraction. Relationship extraction is to judge the semantic relationship between two entities, and extract such triples as (subject, relation, object) from a piece of text. From the definition of relationship extraction, we can also see that relationship extraction mainly does two things: identifying subjects and objects in text (entity recognition task); Determine which relationship these two entities belong to (relationship classification). Relationship extraction can provide support for the automatic construction of knowledge map, search engine, question and answer and other downstream tasks.

(4) Event extraction. The task of event extraction aims to identify specific types of events and present the elements that play a given role in the event in a structured form, which is widely used in the fields of automatic summarization, automatic question answering, information retrieval and so on. The event extraction task can be further divided into four sub tasks: trigger word recognition, event type classification, event argument recognition and role classification. Among them, trigger word recognition is to recognize the core words that promote the occurrence of events in the sentence. Event type classification aims to judge the event type corresponding to the trigger word in the sentence, including birth, marriage, death, etc. Event argument recognition is to identify the argument in the event, that is, the participants of the event. Role classification is to judge the role relationship between trigger words and entities in sentences, such as attackers, victims, etc

Focusing on above four main tasks, in this paper, we introduce the current main research progress of information extraction through detailed literature analysis. Specifically, we first introduce the representative methods in the field of entity extraction, anaphora resolution, relationship extraction and event extraction, and summarize the advantages and disadvantages of these methods. Secondly, we summarize the problems existing in the field of information extraction, and analyze and prospect the future development trend of the field of information extraction.

2. Related concepts of information extraction

2.1. Information extraction

Information extraction extracts the relationship between named entities from the corpus composed of natural language. It is a deeper research based on named entity recognition [1].

In recent years, this work gradually relies more on an algorithm called machine learning algorithm, so machine learning breakthroughs in some aspects provide technical support for information extraction. Golshan proposed that the latest methods in this field include machine learning based methods and depth learning based methods [2]. In recent years, depth learning based methods have been the focus of researchers. All these methods have laid the foundation for the emergence of information extraction technology (IE) [3]. Cui et al. Proposed a neural open IE method [4]. This method is based on coding and framework. It transforms this ie into another problem about sequence. Its characteristic is that the input sequence is a sentence and the tuple with placeholder is its output sequence. Research shows that the performance of this system is significantly better than that of the same type, and its accuracy and recall rate are also significantly higher than other systems.

2.2. Lexical analysis

Lexical analysis is the basis of understanding words, the smallest grammatical unit in natural language. Language is based on words, and words are composed of morphemes, that is, morphemes are the smallest meaningful units of words. Lexical analysis includes two tasks: first, we should be able to correctly segment a series of continuous characters into words one by one; Second: we should be able to correctly judge the part of speech of each word, so as to facilitate the implementation of subsequent syntactic analysis [5]. The correctness and accuracy of the above two aspects of processing will have a decisive impact on the subsequent syntactic analysis, and ultimately determine the correctness of language understanding.

Different languages have different requirements for lexical analysis. For example, there are great differences between English and Chinese in lexical analysis and processing. In English language, the segmentation of words is very difficult, which requires not only the knowledge of word formation, but also the possible segmentation ambiguity. Because English words have complicated changes such as part of speech, number, tense, derivation, deformation and so on, plus English words often have a variety of interpretations, the judgment of word meaning is very difficult. Every word in Chinese is a morpheme, so it is quite easy to find morphemes. It can be seen that in the lexical analysis and processing of natural language understanding, the difficulty of lexical analysis of Chinese, Japanese, Korean and other languages is word segmentation, while the difficulty of English, French and other languages is morpheme differentiation.

2.3. Chinese parsing

Parsing is developed on the basis of computer system. Common applications of parsing include: computer translation, text annotation, one-to-one question answering system, natural extraction of information, automatic search and so on. If you don't know the word parsing, you must know grammatical analysis. These are two different definitions. Parsing, to put it bluntly, is to automatically recognize sentences and syntactic units in a certain rule of grammar, and output recognition according to the regulations.

In the early development stage of Chinese sentence pattern analysis, it basically relies on the way based on basic rules, which adopts the form of unambiguous rules and the behavioral characteristics of interpreting ambiguity. It is based on Chinese Linguistics and emphasizes the process of linguists' understanding of language phenomena through this calculation method. With the development of economy, the calculation method of computer language database is also gradually updated, and the sentence calculation based on basic rules is gradually replaced by a new calculation method - the method based on statistics [6]. In the model of statistical sentence patterns, the form is concise and clear, the utilization rate of parameter space is high, and the efficiency of job analysis is much higher

than that of the previous ones. However, in the context analysis of the article, the sentence pattern analysis method based on statistics is still insufficient. Therefore, on this basis, we need to increase the probability model of structural information and the model based on history. After the methods based on statistics and basic rules, shallow syntax analysis is introduced. Shallow syntactic analysts classify tasks, which are generally composed of two subtasks, namely, the identification of sentence blocks and the attachment between sentence blocks. By solving these two sub problems, and then merge the problems. So as to complete the purposeful parsing task [7].

Chinese language and culture are broad and profound, and many words have multiple meanings. Even a very simple sentence may have many results, so the analysis of sentence grammar is very important. Chinese search volume is too large. Syntax analysis is an extremely complex process. Generally, completing a syntax analysis requires a lot of data support, and there will be different data requirements according to different sentence lengths.

3. Main tasks of information extraction

3.1. Named entity recognition

Named entity recognition is considered by the public and researchers as the most basic work in the field of information extraction. It can recognize proper nouns and related phrases in large paragraphs of the text, but they should be meaningful, and classify and map entities for other subsequent processing of information extraction. Named entity recognition methods can be divided into three categories: dictionary based, rule-based and machine learning based. The dictionary based method uses the existing data resources to locate the named entities that appear in the text. Generally, the accuracy and recall of dictionary based methods are relatively low. The rule-based method recognizes named entities in text by formulating rules related to various features. At present, the most commonly used named entity recognition technology mainly adopts the method based on machine learning.

The early NER method mainly used the rule template manually constructed by linguistic experts according to the characteristics of language knowledge, and realized the recognition of named entities by matching. For different data sets, it is usually necessary to construct specific rules, which are generally constructed according to specific statistical information, punctuation, keywords, indicators and direction words, location words, head words and other characteristics. Krupka et al. Proposed an SRA system for English NER. The system includes two subsystems NameTag and HASTEN. Hasten constructs and generates person name and place name rule templates according to the semantic information of the text for further recognition. Shaalan et al. Used the context features of the text to construct rules, and at the same time added a place name dictionary to identify professional nouns. For the Chinese ner, the initial research focused on the study of professional nouns. Zhang Xiaoheng further summarized the rules according to the structural rules and morphological markers of institution names, and identified university name entities from more than 6 million corpora in three places, with a accuracy rate of 97.3%. From the perspective of professional noun recognition, Wang Ning and others fully considered the characteristics of the financial field, and used rule-based methods to study the recognition of company names. This method analyzes and studies the financial news text, summarizes the structural characteristics and context information of the company name, summarizes and forms a knowledge base, and adopts the strategy of two scans for recognition.

The method based on deep learning is similar to the processing flow of NER and other sequence annotation tasks. First, the sequence is converted into a distributed representation by encoding methods such as Word2Vec, then the feature representation of the sentence is input into the encoder, and the feature is automatically extracted by neural network. Finally, the decoder such as CRF is used to predict the tags corresponding to the words in the sequence. In the early stage, researchers mostly conducted in-depth research on ner methods based on supervised and remote supervised deep learning. Since the pre training model Bert (bidirectional encoder representation from transformers) was proposed in 2018, it has also attracted researchers' attention. Recently, the method based on cue learning has also been preliminarily tried on ner task and achieved success. A neural network method

is based on deep learning, and its general process is divided into four steps: ① Sequence: input sequence after preprocessing. ② Word embedding: converts an input sequence into a vector representation of a fixed length. ③ Context encoder: embed words for semantic coding. ④ Tag decoder: further tag decoding.

The NER method based on HMM uses Viterbi algorithm to allocate the possible target sequence to each word sequence, which can capture the locality of the phenomenon and improve the performance of entity recognition. Based on the characteristics of case, number symbols, sentence first words and so on, Bikel and others use HMM to calculate the probability that a word is an entity type. However, the model still cannot capture long-distance information, and there are still some unrecognized entities. Zhou et al. Proposed a new method of chunk marker that is based on HMM, which expanded the internal semantic features, internal place name dictionary features and external context features on the basis of Bikel, and improved the traditional formula of HMM, so that more context information can be fused to determine the current prediction type. For Chinese NER, Zhang Huaping proposed an automatic Chinese name recognition method based on role annotation with the help of HMM. This party adopts HMM to mark the role of the word segmentation results, and identifies and classifies named entities through the maximum matching of the best role sequence. This method solves the problems of the loss of names without obvious features, and the difficulty of recalling names with internal words and context words. Yu Hongkui et al. Proposed a Chinese ner model based on cascaded HMM, which is composed of three-level HMM. After word segmentation, the low-level HMM identifies common non nested names of people, places and institutions, while the high-level HMM identifies nested names of people, places and institutions.

3.2. Anaphora resolution

It mainly describes the correlation between text concepts, which is usually divided into anaphora and co reference. The current anaphora and the association between words, phrases or some sentences refer to the close semantic aspect, which is called anaphora, while CO anaphora refers to two or more nouns or noun phrases pointing to the same reference, and anaphora does not depend on the specific context. Anaphora resolution is actually a process of establishing the relationship between concepts, which is one of the core issues of text processing.

Early researches on anaphora resolution used a large number of manually constructed domain and language knowledge to form logical rules for resolution. Examples of representative work are as follows

Traversal algorithm based on complete analytic tree

In 1978, Hobbs proposed a anaphora resolution algorithm that does not rely on any semantic knowledge or text information, but only uses grammatical rules and complete parsing tree information. The algorithm first establishes a complete parsing tree for each sentence in the document, then traverses the complete parsing tree using the left to right breadth first search method, and finally selects the legal noun phrase as the antecedent according to the dominance and binding relationship in the grammatical structure. This algorithm is proposed in the form of model and is rarely used directly in practical systems.

3.3. Relation extraction

One of the more important research topics of IE is a function of relationship extraction, which is used to obtain the grammatical relationships between entities, which can also be semantic relationships. In daily applications, identifying entities in text is the first step of information extraction, and the more important is to determine the relationship between entities. Similar to named entity recognition, the types of entity relationships are pre-defined, such as geographical relationships between cities, people and organization relationships. Usually, people regard the problem of relationship extraction as a classification problem, and initially use the method based on knowledge base to extract relationships.

Supervised relation extraction has been widely recognized as a method with high accuracy by people in relevant industries. According to the manual annotation data classification and training model that people have mastered at this stage, this thinking mode is used by this method, and then start the next step of some extraction and matching recognition of specific relationships. The method based on feature vector mainly constructs feature vector by extracting feature information such as syntax and grammar from sentence context, and then trains entity relationship recognition model by using the similarity of feature vector to complete entity relationship recognition and extraction. Kamath et al. uses the maximum entropy classifier to build an extraction model. By adding text features, it can use few vocabulary features to achieve good results, so as to achieve an effect that is to reduce the dependence on the extraction tree. Based on the evaluation materials of SEM eval-2007, through the characteristics of entity context, distance and so on, with the help of support vector machine (SVM) model, the extraction effect with F value of 71.8% is achieved; Tratz et al. Used the maximum entropy classification method to achieve the extraction effect on the SEM eval-2010 evaluation corpus, and the F value is 77.57%; Culotta et al. Used the classification method of conditional random fields to extract relationships with the help of Mallet CRF with default regularization parameters, and achieved good results.

3.4. Event extraction

In information extraction, an event refers to an event that occurs in a specific time period and region, is participated by one or more roles, and is composed of one or more actions. Event extraction mainly studies how to extract the events that users are interested in from unstructured text, and describe these events in the form of structured text for further query, tracking and analysis.

The information covered by the subject event can also be described. These can be distributed in a document, and of course, they can also be in some documents. The extraction method used by meta events is only in some sentences. How to determine a document set that can describe the same theme events is the key to theme event extraction, and merging some centralized and scattered Theme Event fragments through intra text or cross text technology is also one of the keys. In research, event framework or ontology is usually used to represent the basic components of the subject event and the relationship between the components.

This method of extracting theme events based on event framework mainly uses the defined structure and an event framework, that is, hierarchical to extract, and the use of this framework can be used to summarize and express events. Framework is a common knowledge representation method, which can be used to describe the contour framework of related concepts. M. Minsky put forward the concept of framework in the article "a framework for retaining knowledge", which attracted the attention of scholars. In the face of new situations, people will look for the existing models in their brain to help them understand a new thing. These typical situations are knowledge frameworks. Just like meeting activities, people can associate from the time, place, participants, influence and organization of the meeting. The sides of an event can be semantically separated, so the framework here is actually a classification system used to separate the different sides involved in an event. The words used to describe different aspects of events are "side words" of events, and the event framework is a classification system composed of "side words".

Cheng et al. Introduced a fuzzy event extraction agent system based on ontology. In the construction of ontology, a four layer ontology construction model is proposed, including domain layer, category layer, event layer and extended concept layer. Domain represents the domain name of the ontology and is composed of several categories defined by domain experts; Each category contains a set of events; The event layer defines that each category contains event types; The extended concept layer includes event concepts and object concepts, and defines the roles and concepts corresponding to each type of event and the corresponding sub events. The ontology constructed by the model is applied to the extraction of news events and automatic summarization. Experiments show that the system can better extract Chinese Meteorological news events.

4. Web information extraction algorithm

4.1. Concept of Web information extraction system

Combined with various descriptions of the concept of information extraction and the discussion of information extraction technology in a series of message understanding conferences (MUC) in the past 20 years, take a comprehensive view of the definitions, The concept of Web information extraction can be defined as: Web information extraction (WIE) is the process of extracting a specified type of information (events, facts) from web page text and filling it into a structured data base for users to query.

The core of Web information extraction technology is to identify the data that users are interested in from the unstructured or semi-structured information contained in web pages, and transform it into a more structured and semantically clear format.

4.2. Algorithm flow of dynamic web page information extraction based on sequence comparison

The web information extraction method proposed in this paper is divided into three steps:

(1)Serialization. Convert the web pages in the training set into string sequences.

(2)Common framework detection. Sequence alignment algorithm is used to compare string sequences, and divide the common framework and data domain. Among them, the public framework refers to the parts that are common among web pages such as header information, tail information, advertising, browsing guide bar and flash and have nothing to do with the substantive content of web pages. Data field refers to the content unique to each web page except the public framework. Then, the string sequence of the data field is transformed into a label tree structure to obtain the sample set of step (3).

(3)Template extraction. Search and divide the matching and mismatching parts between samples, and get the template after regularization.

5. Discussion

In the continuous construction of corpus, the main development trend of natural language processing has become large-scale language and data processing. At the same time, statistical mathematics methods have been paid more and more attention, and the methods of machine automatic learning to obtain language knowledge in natural language processing are also more and more extensive. So far, the research field of natural language processing has been expanding, from text to speech intelligent recognition, syntax analysis, a series of machine learning, translation and information retrieval. Natural language processing is also constantly providing assistance for the development of other emerging disciplines or fields, such as biological information. The future research focus also has the goal of how to improve the ability of computer to process language.

The research of natural language processing is somewhat abstract, but its basic research lies in the research of grammar and syntax, and its focus is on text content and language. Understanding language requires the use of logic, but also requires a lot of knowledge reserves. These two points are the difficulties of natural language processing. These supports are needed to better process data and further understand and analyze texts.

In the long run, natural language processing can be applied in many fields and will have a good development. It is a newly emerging scientific field, which can have a great impact on many disciplines. In the future, its development is likely to change from manual construction to automation. Explicit knowledge can be used by people. In order to explore the relationship between the components of language, so as to avoid the tedious and time-consuming manual work. In the level of text understanding and reasoning, we can complete the deep understanding of the text from shallow to deep.

Natural language processing is a rapidly developing and emerging discipline. Its growth process includes imaginable challenges and difficulties. Related systems in the field of natural language processing, such as machine translation and speech recognition, do have shortcomings, but natural language processing is widely used, so these can prove its important position in the field of science and technology. It is believed that NLP will have a brighter future in the near future.

6. Conclusion

Information extraction has become a research hotspot in the field of natural language processing. This paper first reviews the development of information extraction. Secondly, the research progress of key technologies of information extraction is summarized from four aspects: named entity recognition, anaphora resolution, relationship extraction and event extraction. Finally, we analyze some main problems of information extraction and predict the research trend of information extraction.

References

- [1] Guo X Y, He T T, Hu X H, Chen Q J Chinese entity relation extraction based on syntactic semantic features [J]. Chinese Journal of information technology, 2014,28 (06): 183-189.
- [2] Golshan P N, Dashti H A R, Azizi S, et al.A Study of Recent Contributions on Information Extraction[J].2018.J.Piskorski and R.Yangarber.Information extraction.
- [3] J.Piskorski and R.Yangarber,“Information extraction:past,present and future,”in Multi-source,multilingual information extraction and summarization,Springer,pp.23-49,2013.
- [4] Cui L,Wei F,Zhou M.Neural Open Information Extraction[J].2018.
- [5] Liu Q, Jia H B Research and Prospect of automatic word segmentation technology in Chinese information processing [J]. Computer engineering and application, 2006, (03): 175-177+182.
- [6] Yuan L C. Syntax analysis method based on statistics [J]. Journal of Central South University (NATURAL SCIENCE EDITION), 2014, 08:2669-2675.
- [7] Zheng W F. A review of Chinese parsing research [J].Information technology, 2012, 07:72-74+78.
- [8] Feng Z W. The history and current situation of natural language processing [J]. China foreign languages, 2008, 5 (1): 14-22.
- [9] Wang T, Mai F J, Liu Z. Research on natural language processing and its application prospects [J]. Journal of Guilin Institute of aerospace technology, 2006, 11 (4): 19-21.
- [10] Zong C Q Statistical natural language processing [M]. Beijing: Tsinghua University Press, 2013:486719726.
- [11] Niklaus C,Cetto M,Freitas,André,et al.A Survey on Open Information Extraction[J].2018.
- [12] Grishman R, Sundheim B.message Understanding Conference on Computational Linguistics COLING—96, 1996—08.
- [13] Soderland.Learning Information Extraction Rules for Semistructured and Free Text.Machine Learning 1999.
- [14] Liu S Q, Zhu M, Tan X B A web page semantic block mining algorithm based on tree matching [J]. Small microcomputer system, 2009, 30 (8): 1541-1545.
- [15] Han Z M, Li W Z, Mo Q Research on effective HTML text information extraction methods [J]. Research on computer application, 2008, 25 (12): 3568-3574.
- [16] Zhang S X Research on key technologies in information extraction [D]. Beijing: Beijing University of Posts and telecommunications, 2007.
- [17] Valter C, Giansalvatore M, Grammars have exceptions, Inf.Syst.23(8)(1998)539–565.

- [18] Joachim H, Jason M H, Hector G M, Semistructured data:the TSIMMIS experience, in:Advances in Databases and Information Systems, 1997, pp.1-8.
- [19] Arnaud S, Fabien A, Building intelligent web applications using lightweight wrappers, Data Knowl.Eng.36(3)(2001)283-316.
- [20] Deng C, Shipeng Y, Ji-Rong W, et al.VIPS:a vi-sion-based page segmentation algorithm[R].USA:Microsoft Technical Report, 2003.