Research on Progress and Inspiration of Entity Relation Extraction in English Open Domain

Xu Jian[1,2], Yao Xianming[2(✉)], Gan Jianhou[1], and Sun Yu[3]

[1] Key Laboratory of Educational Informatization for Nationalities (YNNU), Kunming 650500, China
qjncxj@126.com
[2] School of Information Engineering, QuJing Normal University, Qujing 655011, China
yxml76@qq.com
[3] School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China
sunyu_km@hotmail.com

Abstract. In the era of big data, how to extract unrestricted type of entity relations from open domain text is a challenging topic. In order to further understand related deep issues, this paper summarized the latest progress in the field of English entity relation extraction, ranging from binary to n-ary entity relation extraction; furthermore, some milestone systems are introduced in detail. This paper makes a preliminary exploration on the extraction of entity relations in the Chinese open domain. In particular, the inspiration of English to Chinese has also promoted the development of Chinese entity relation extraction.

Keywords: English open domain · Binary entity relation · N-ary entity relation · Entity relation extraction

1 Preface

With the advent of the era of big data, traditional entity relation extraction algorithms are faced with a series of problems such as weak domain expansion capabilities, restricted entity types, and heavy manual labor. Therefore, the extraction of open domain entity relations has become a hot spot.

The research of English open domain entity relation extraction has started earlier, and gone through the process from binary to n-ary entity relation extraction. Lots of milestone systems such as Reverb [1], OLLIE [2], Kraken [3], ClausIE [4] and so on emerged. The relation extraction in the open domain of English is becoming more and more mature. Nevertheless, relevant research in the field of Chinese has just begun, and there is still much room for exploration.

In order to promote the development of related research in the field of Chinese, this paper makes a review of English open domain entity relation extraction, ranging from binary relation to n-ary entity relation. Several milestone systems are presented in detail as typical cases. Furthermore, this paper also makes a preliminary exploration of entity

X. Chen et al. (Eds.): ML4CS 2020, LNCS 12487, pp. 42–55, 2020.
https://doi.org/10.1007/978-3-030-62460-6_5

relation extraction in Chinese open domain, and analyzes the Inspiration of English to Chinese. Considering the lack of relevant research in the field of Chinese, we hope this paper can contribute to the current work.

2 Summaries

In order to have a comprehensive understanding of open domain entity relation extraction, we would like to introduce its definition, tasks and applications in this section.

2.1 Definition

Entity relation extraction refers to extract semantic relations between entities and expressions from text. This kind of semantic relations could reflect interactions between them [5]. Currently, there is no authoritative definition to open domain entity relation extraction. Definition defined in paper [6] is generally accepted: Open information extraction is a novel extraction paradigm that tackles an unbounded number of relations. It usually refers to extending the traditional entity relation extraction from specific domain to general domain. For example, in the sentence of "Einstein was born in Ulm, Germany", there exists a relation of (Einstein, be born in, Ulm Germany).

Let E be a set of entities in text, e_i and e_j $(0 < i, j < |E|)$ are elements in E, and let R be a relation set in the same text, $r_k (0 < k < |R|)$ is an element of R, so that I is a set of random combination of e_i, e_j, r_k in the form of (e_i, r_k, e_j), and the task of entity relation extraction is to extract $\delta = < \mathrm{E},\ R, I >$.

In the triple of (e_i, r_k, e_j), r_k is usually called as relation or relation phrase consisting of one or more words. In the task of extracting relations delivered by predicate, r_k is a predicate or P for short. e_i and e_j are often called entities, appropriately, the relation between them is expressed as r_k. In some studies, e_i and e_j are named as arguments of r_k. Also, in the task of extracting relation delivered by predicate, e_i and e_j act as subject and object respectively; and this kind of relation could be expressed as (Subject, Predicate, Object), or (S, P, O) for short.

2.2 Tasks

The task of entity relation extraction is to extract the semantic relation between entities represented by a continuous word. Based on the analysis of previous studies, this paper divides these relations into three categories.

- **Relations delivered by verbs.** Since verbs between subject and object could clearly reflect their interaction in sentence, verbs are a good choice for relation extraction. For example, in sentence of "*Bin Laden masterminded the 9/11 attacks*", the verb of "*masterminded*" could reflect interaction of "*Bin Laden*" and "*9/11 attacks*", so that it could be selected as the relation of them. In systems such as Reverb [1], WOE [6] and TextRunner [7], verb phrases are the main target for extracting target in the domain of English, as is the domain of Chinese.

- **Relations delivered by common nouns.** In natural language, there is a structure in the form of "*proper noun + common noun + proper noun*", in which *proper noun* is noun or adjective of location, people, organization, etc. *Common noun* is general concept which can be considered as attribute. In this kind of structure, triples of (Entity, Attribute, Value) could be mined as entity relation. For example, in the sentence of "*American President Obama is going to cooperate with their western allies*", there is a fragment of "*American President Obama*" which satisfies this kind of structure, and we could find an entity relation of (*America*, *President*, *Obama*). However, due to inversed order, long distance dependency, and instable structure, extraction of this kind of entity relation is rarely mentioned except for OLLIE in paper [2].
- **Relations established contextually.** With the passage of time, people's understanding of objective laws will also change, which means that certain triples extracted from previous rules will be invalid, so a certain dependence should be given. Therefore, extracting relations established contextually is particularly important for open domain entity relation extraction. For example, we could extract relation of (*the earth; be the center of; the universe*) from the sentence of "*Early astronomers believed that the earth is the center of the universe.*" However, we know that this relation is invalid, and the valid relation should be (*(the earth; be the center of; the universe) AttributedTo believe; Early astronomers*). The OLLIE system extends the representation of open information extraction and allows it to accept other contextual information, such as attributes and clause modifiers, thereby solving this type of problem. Systems such as TIE [9] and Yago2 [10] extended temporal constraints and achieved certain success in given types of entity relation extraction. Yao uses relation nested in subjects and objects to extract compound entity relations, thus resolving this problem. For example, they could extract the relation of (*Early astronomers, believed, (the earth, is the center of, the universe)*) from above example.

Of course, not only the above relations exist, but with the development of research, more types of entity relations will be discovered in the future. In addition, there are other types of relations in other studies, but this paper has not been found. Yang Bo [12] proposed implicit entity relation extraction, which is different from the method of extracting explicit entity relations in this paper, so it is not taken into account.

2.3 Application

Because the open domain entity relation extraction can expand the types of entities and relations, and also has the domain extension function, it has been applied in many applications and played an important role.

- **Knowledge base construction** [13, 14]**.** Generally speaking, knowledge base consists of domain ontology and knowledge graph. It is composed of vertexes and edges. Vertexes refer to concepts and entities, and edges refer to relations between concepts and entities in knowledge base. This kind of structure is similar to triple discussed in this paper, so entity relation extraction is especially important to the construction of knowledge base. Prior studies on the construction of knowledge

base mainly focus on a given domain, thereby limiting the types of entities and relations. By the introduction of open domain entity relation extraction, these problems could be resolved perfectly. Nowadays, lots of knowledge graphs have appeared, such as Google Knowledge Graph, Satori of Microsoft, etc. There are also some general knowledge graphs in the Chinese field, such Baidu ZhiXin, Sogou ZhiLiFang, and CN-DBpedia developed by Knowledge Works Research Laboratory in Fudan University. The emergence of these knowledge graphs has greatly promoted the development of in-depth question answering, intelligent information retrieval and knowledge reasoning.

- **Semantic Search** [15]**.** Traditional search engine works in a workflow of analyzing question raised by user, extracting key words as well as computing similarity between key words and web pages, relevant pages with higher similarity will be provided to user. This kind of information retrieval can only match at the word level, and cannot understand users' purpose, so relevant pages cannot meet users' needs. If the task of entity relation extraction is introduced into traditional information retrieval to capture the relation between entities and perform semantic level information retrieval, it will increase the probability of information retrieval to find more closely related pages and even give direct answers to improve user-friendliness.
- **Question Answering System** [16, 17]**.** This is nearly the same problem with information retrieval, for entity relation extraction could catch semantic relation between entities, these information could be provided to question answering system for modeling user's attention, deep analysis to sentences will help positioning answer. In addition, for the uncertainty of answer types, any kind of entity relation could be treated as candidate answer, extracting relation from answer nested sentences will be helpful to the generation of answer.

Generally, as one of the most important tasks in information extraction, entity relation extraction not only plays an important role in the above three applications, but also has a wide range of applications in natural language processing, machine translation and machine reading. Due to space limitations, these applications will not be discussed here.

3 Summaries on English Open Doman Entity Relation Extraction

Since the relevant research in the field of English is earlier than Chinese and has accumulated a lot of experience, these studies have good reference value for other languages. Here we will introduce some excellent research. According to the summary of some domestic scholars, the open domain entity relation extraction in English can be divided into two stages: binary entity relation extraction stage and n-ary entity relation extraction attachment. Next, we will introduce some milestone systems based on this summary. The presentation includes tasks, techniques, experiments and deficiencies.

3.1 Binary Entity Relation Extractions

Binary entity relation extraction is the earliest research work, and has the most fruitful work and the most mature technical solutions in the open English domain. Extensive research has been conducted on the extraction of relation phrases, relation arguments, conditional dependence, juxtaposition and other issues. Data size and speed as well as the precision and recall rate have been improved greatly. In the meantime, it also provides technical premise to n-ary entity relation extraction. The research results at this stage mainly have six systems, namely, KnowItAll [18], TextRunner [7], WOE [6], Reverb [1], R2A2 [19] and OLLIE [2].

KnowItAll

KnowItAll system is an advantageous attempt to extract information from restricted domain to open domain. Its main purpose is to solve the problem of obtaining concept examples in an open domain environment. You can give concept name to be queried, the system can use a search engine to retrieve the text in the Web and return its instance.

KnowItAll first constructs a baseline system based on Hearst patterns [20], and instances of these patterns will be provided to search engine to retrieve more instances which will be added to knowledge base. KnowItAll also proposes the tasks of Pattern Learning, Subclass Extraction and List Extraction. Task of Pattern Learning is to learn and evaluate domain specific patterns automatically according to Hearst which belong to open domain. Task of Subclass Extraction can learn subclass of a given concept so as to construct concept tree. Task of List Extraction is the population of concept with instances according to structural similarity of web pages.

TextRunner

TextRunner can be regarded as the first generation open domain entity relation extraction system [21]. This is the first work to incorporate remote supervision into the research of entity relation extraction. TextRunner transforms the relation extraction problem into a classification problem, aiming to identify relation phrases to achieve its goal.

Rule-based patterns are adopted to automatically tag training instances. Tagged entities mainly include noun phrases, and relation phrases are composed of continuous words in syntactic structure between entities. Classification features include universal features such as POS, length of relation phrase as well as the number of stop words in relation etc. Deep syntactic features are not in consideration so as to speed up system performance. Finally, naïve Bayes was borrowed to train classification model.

WOE

Alike to TextRunner, WOE (Wikipedia-based Open Extractor) adopted distant supervision to automatically tagging training instances. After extracting effective features, instances will be sent to train classification model. Unlike TextRunner which performs the task of tagging based on hand crafted rules, Wikipedia inforbox information was selected as tagging data source. For the high quality of tagging data, classification model can outperform TextRunner.

WOE trained two kinds of systems: WOE^{pos} and WOE^{parse}. Both of them realized the same purpose of extracting entity relation, but WOE^{pos} uses shallow syntactic features which is same to TextRunner, while WOE^{parse} uses deep syntactic features both in training and recognizing. Probability model and conditional random field model were chosen as training model respectively.

The performance of two systems is as expected. The performance of WOE^{pos} is almost the same as TextRunner, but lower than WOE^{parse}. In consideration of speed, WOE^{pos} is also the same as TextRunner, and far faster than WOE^{parse} which is attributed to deep syntactic analysis. The comparison of two systems shows that deep syntactic features are not fit for open domain entity relation extraction, since they are time-consumed, but the speed is especially important to open domain.

Reverb

Previous studies like TextRunner generically take continuous words between entities as relation phrase which may take incoherent and uninformative extractions. The greatest contribution of Reverb is to find the boundary of relation phrase. Reverb introduced LVCs (light verb constructions) theory, furthermore, lexical and syntactic constraints were defined to extract relation phrase.

Syntactic constraint could reduce incoherent and uninformative extractions errors. Lexical constraint could reduce wrong relation phrases with lowest occurrence which cannot be filtered by syntactic constraint. At the stage of implementation, syntactic constraints were encoded as regular expressions to match corresponding relation phrase in sentence; entities were extracted from left and right side of relation phrase to form a triple. In the following, over-specified relation phrases will be filtered by threshold, defined by lexical constraints. Finally, linear regression classifier was adopted to assign a confidence score to each triple, aiming to find a balance between precision and recall rate.

R2A2

Entity relation extraction should focus on three kinds of extraction tasks as triple listed: first entity, relation phrase and second entity. As far as the relation phrase extraction task is concerned, Reverb has provided technical solution. Previous studies had no deep studies in entity extraction. R2A2 (REVERB relation phrases with ARGLEARNER's arguments) is just this kind of system, which focuses on extracting entity that makes open domain entity relation more perfect.

The analysis of R2A2 to real text shows that the left and right position for first entity and second entity could be divided into several categories which enable supervised method to be borrowed for recognizing the boundary of each entity. R2A2 designed 3 classifiers. The first classifier is the left boundary recognizer for first argument (R2A2 names entity as argument, which means argument for relation). The second classifier is right boundary recognizer for first argument. The third classifier is right boundary recognizer for second argument. Note that there is no left boundary recognizer for second argument, and statistics show that almost all the left boundary of second argument is just at the end of relation phrase. Contextual features, length of sentence, POS, case and punctuation were selected as features. Training instances come from sematic role labeling. Relation phrases come from the result of Reverb. CRF, REPTree and CRF were chosen as training models for each classifier respectively.

Evaluation shows that the precision of first argument increased by 10%, and 20% for second argument which is better than previous systems. It also outperforms previous systems in other respects.

OLLIE

OLLIE (Open Language Learning for Information Extraction) is a more perfect system in the field of open domain entity relation extraction. As introduced above, systems such as Reverb and R2A2 had made great contributions to the recognition of relation phrase and entity (argument). However, there are still two problems. The first is that relation phrases delivered by noun were hardly mentioned, while the second problem is that contextually established relations were rarely mentioned. Aimed at resolving these problems, OLLIE promoted two methods: (1) expanding the syntactic scope of relation phrases to cover a much larger number of relation expressions, and (2) expanding the Open IE representation to allow additional context information such as attribution and clausal modifiers.

In the implementation stage, OLLIE involves three steps. First, seeds of entity relation with high quality provided by Reverb will be created to bootstrap a very large training set which encapsulates the multitudes of ways in which information is expressed as text. Second, open pattern templates will be learned by identifying dependency path between relation phrase and corresponding arguments. Finally, extracting new entity relations using learned pattern templates.

Experiments show that the correct extraction rate of OLLIE is 4.4 times higher than REVERB, 4.8 times higher than WOE^{parse}, and the accuracy is about 75%.

3.2 N-ary Entity Relation Extractions

After a series of studies, there may be a lot of research space in binary entity relation extraction. It has made great progress, making it difficult to improve accuracy and recall rate. Therefore, a lot of work has turned to n-ary entity relation extraction. Statistics in paper [22] show that n-ary entity relations occupy about 40% in all entity relations, and n-ary entity relation could make the extracted result easier to be understood. So the n-ary entity relation is a novel and fruitful direction.

TIE

Although n-ary entity relation extraction is pretty important, and relevant studies started at a later time, resulting in fewer achievements. TIE (Temporal Information Extractor) is one of such system which extracts n-ary entity relations. TIE believes that most entity relations have temporal constraint with start and end time, and temporal constraint will make relation meaningful. Temporal information was extracted from the text as an argument of entity relation. Temporal entropy was adopted to assess the performance of TIE. Experiments and its wide application had proved its usefulness. TIE is one of these systems that raised n-ary entity relation task, and certain progress has opened up a new direction for entity relation extraction.

Yago2

Yago2 is a relatively ripe knowledge base whose data comes from Wikipedia. It owns declarative rules including factual rules, implication rules, replacement rules and

extraction rules, so the number of entities in its base increased greatly. Temporal information, space information and contextual information are also extracted as arguments of entities. Temporal information contains begin and end time of an entity. Space information includes longitude and latitude. Contextual information refers to textual context that entity occurred in Wikipedia pages. In this way, an entity relation with six parameters including predicate (relational phrase) is constructed, namely subject, predicate, object, temporal, space and context, which is called as SPOTLX for short. So Yago2 can trace the process of all entities and events in knowledge base from birth to death, including its specific geographic location information. Finally, Yago2 integrates information such as entities and events into the map, allowing users to query corresponding information in the form of visualization; therefore, the data is more intuitive.

The limitation of Yago2 is that it only considers time, space and environment information. The domain it focuses on has certain limitations, and it is difficult to expand to more constraints of entity relation arguments.

Kraken

In contrast, Kraken (N-ary OIE fact extraction system for facts of arbitrary arity) system [3] is a relatively more mature system in n-ary entity relation extraction. Based on dependency analysis, it can extract n-ary entity relation by locating the relation phrases and obtaining their corresponding arguments. The specific algorithm steps are as follows:

1. The detection of fact phrase. Detecting event relation phrase composed by verbs, modifiers and prepositions, such as “has been known”, “claims to be”. If there is only one predicate, it can be also considered as fact phrase.
2. To find argument heads. Finding subject heads for each fact phrase according to dependency path like “nsubj-↓”, “nsubjpass-↓, rcmod-↑”, “appos-↑” etc. Finding object heads according to dependency path like “dobj-↓”, “prep-↓, pobj-↓” etc.
3. Detection of full arguments. Follow all downward links recursively from the argument head in order to get the full argument, excluding any links that were part of the type-path to the argument head.

Experiment shows that the precision of Kraken is 68% and the completeness is 79%, while the highest precision of Reverb is only 64% and the completeness is only 44%, which reflects the excellent performance of Kraken system. But in terms of extraction speed, Kraken consumes 319 ms for 500 sentences; Reverb only spends 13 ms, with a speed difference of nearly 20 times, which fully proves that deep syntactic analysis will have a greater impact on extraction speed.

ClausIE

ClausIE (Clause-based open Information Extraction) [4] is a relatively more mature and powerful open information extraction system, which can consider unary, binary, ternary and n-ary entity relation.

ClausIE uses linguistic knowledge of English grammar to detect clauses, and then recognizes the types of clauses according to grammatical functions of each clause component. It listed seven types of clauses: ① SVi; ② SVeA; ③ SVcC; ④ SVmtO; ⑤ SVdtOiO; ⑥ SVctOA; ⑦ SVctOC. For any given clause, ClausIE can determine

the type of the clause by its predicate, and obtain the information of the clause from the dependency analysis results as well as the type of predicate in domain independent dictionary.

In the phase of system evaluation, they compared ClausIE to TextRunner, Reverb, WOE, OLLIE and Kraken. Experimental datasets include Reverb dataset, Wikipedia and New York Times. Statistics shows that ClausIE could extract more entity relation instances than other systems with equal or even higher recall and precision rate. Although OLLIE is better than other previous systems, the correct instance extracted by ClausIE is 2.5–3.5 times that of OLLIE. Errors of ClausIE are caused by the result of the wrongly analyzed dependency.

3.3 Reflections and Prospect of English Open Domain Entity Relation Extraction

Although the entity relation extraction in English open domain has begun very early, and has achieved many achievements and mature technical solutions, there are still some problems, which will have a certain impact on other research.

- **Lack of standard evaluation dataset and criterion** which would makes it difficult to evaluate the performance of each system horizontally. To evaluate datasets, KnowItAll adopted Tipster Gazetteer and Internet Movie Database, while TextRunner adopted their own data collected from 9 M web pages. WOE adopted Penn Treebank, Wikipedia and the general Web. Reverb used a test set of 500 sentences sampled from the Web, using Yahoo's random link service. OLLIE created a dataset of 300 random sentences from three sources: News, Wikipedia and Biology textbook. These evaluation datasets come from different sources with different quality. The performance of each system may vary greatly, so it cannot reflect the actual performance.
  To evaluate criterion, most of these systems adopted precision and recall rate, while Reverb and OLLIE partly or fully used AUC. As a result, the horizontal comparison between different systems is not very intuitive, and may hinder the development of open domain entity relation extraction indirectly. If we can carry out corresponding academic conferences and competitions, release standard evaluation corpus, and establish a unified evaluation standard, just like what traditional information extraction do, it will undoubtedly promote the development of open domain entity relation extraction.
- **Lack of standard definition of relevant concepts.** Different studies define these concepts from different perspectives. In the definition of binary entity relation extraction, most systems adopted the definition in Sect. 2 of this paper to extract entity relations in the form of (e_i, r_k, e_j) and name it as binary entity relation extraction. However, KnowItAll defines concept instance extraction as unary entity relation extraction, and others as n-ary. In the definition of n-ary entity relation extraction, Kraken aims to extract multiple triples of (e_i, r_k, e_j) from sentence, and this kind of extraction is considered as n-ary entity relation extraction by default which shares the same form with binary entity relation extraction defined in Sect. 2. Nevertheless, this kind of definition neglected their close relation between different

triples extracted from the same sentence. ClausIE aimed at solving this problem and extract more entity relations from sentence. Unlike other studies, entities extracted from sentence are regarded as arguments of relation phrase, so their triples could be unified as an organic whole. But this kind of extraction is obviously different to other works.

- **Introduction of dependency parsing is controversial, but it is a trend.** Due to the long-distance dependence between entities and their relations in text, it can only be solved by dependency parsing. However, dependency parsing may be tedious and time-consuming, which makes it not suitable for processing large-scale data in open domain. In TextRunner and WOE^{parse} systems, they try to obtain syntactic features to improve system performance. Experiment shows that syntactic features do play an important role, but it is time-consuming. In the system of OLLIE, the dependency path is used as template directly, and achieved great success. Later, Kraken also used dependency parsing to extract n-ary entity relation. Although dependency parsing could be time-consuming, scientists are still willing to use it to improve the precision and recall rate, which proves the effectiveness of this technology indirectly. It is believed that with the improvement of computer speed and the development of natural language processing, dependency parsing should be one of the most promising directions in future, or at least achieve a certain balance between speed and performance.
- **Fewer researches reported on entity relation extraction in other languages.** At present, studies are mainly focus on the domain of English, and had achieved a large amount of research results. There are in-depth analysis and technical solutions to problems involved in English. But for other languages, due to various reasons, related studies were rarely mentioned. For example, in the domain of Chinese, although there have been some researches in open domain, it is mainly based on binary, and the research results are relatively fewer, while n-ary is rarely reported.

4 Chinese Open Domain Entity Relation Extraction and Its Inspiration

Compared with English, study of open domain entity relations in other languages is relatively lagging behind, and related results are relatively lacking too, which may be attributed to many factors. As one of the six official languages of the United Nations, Chinese has the most users in the world, which makes Chinese information work of great value. Similarly, it is equally important to do relevant studies in open domain of Chinese. If we can learn from the beneficial experience of English and carry out relevant research work in combination with the characteristics of Chinese, it will bring great impetus to relevant research.

4.1 Introduction of Chinese Open Domain Entity Relation Extraction

Chinese open domain entity relation extraction started at around 2014, and certain achievements have been reached at present, but representative systems have not yet

appeared. This paper divides relevant researches into two categories: one is based on statistics, and the other is based on natural language processing.

Statistics-based methods design corresponding constraint rules according to objective statistical results of entity relations in text, aiming to extract entity relation instances. Qin bing [23] exploited using word distance and entity distance constraints to generate candidate relation triples from the corpus, and then adopted global ranking and domain ranking methods to discover relation words from candidate relation triples, and finally filtered candidate triples by using the extracted relation words and some sentence rules. Result shows that the precision is higher than 80% when extracting large scale relation triples. Yu li [24] adopted approximate method, by making statistics on part of speech, position, distance and other features of the relation words in geographical entity relation, Yu li also constructed an evaluation function to extract relation words, and obtained the accuracy rate of 80% and the recall rate of 87.79%.

Natural language processing based methods makes full use of dependency syntax analysis and utilizes dependency paths as templates or features to extract entity relations. These approaches benefit from the beneficial attempts of OLLIE, Kraken and other systems in English domain and have achieved success in Chinese [13]. Song qing [8] adopted method used by OLLIE system, and select relation with high confidence to extract seed set, matching it in open text, and then save the dependency path of entity relation as templates, and then templates used frequently are selected as seeds to extract new entity relation, so as to realize the automatic extraction of entity relation. Yao xianming [11] uses the idea of the Kraken system as reference, based on the characteristics of the Chinese language; take the verb as the relation indicator to extract its subject and object, and use the nested relation between entity relations to extract contextually dependent entity relations. Liu shen [25] takes dependency relation as the feature of entity relation, and deep learning is utilized to study the entity relation extraction problem. In addition, this paper also studied argument boundary identification problem similar to the R2A2 system in English domain.

In addition, there are other related studies. For example, literature [26–30] presents some different ideas and technical routes, which will not be introduced in-depth here.

4.2 Inspiration of English to Chinese

Through comparative analysis of Chinese and English, we can find that Chinese open domain entity relation extraction is relatively small, scattered, and lacks continuity. These problems make it difficult to formulate a complete technical solution. Once we can learn successful experiences from abroad and apply them to the Chinese field, technology development will be more rapid and beneficial.

- English open domain entity relation extraction experienced a process of relation phrase extraction, entity extraction, attribute extraction and contextual dependence extraction. Different systems have completed part of the tasks and brought them together to form a unified technical solution. Although accomplished by different researchers, these studies are continuous. But for Chinese, relevant studies are isolated, and have no special research to specific problem; therefore, these studies could not be bound together easily. If we can refer to the research ideas in English

and break down tasks of entity relation extraction, we may be able to open a new situation for entity relation extraction in Chinese open domain.

- Extraction techniques in English went through the stages of distant supervision, lexical pattern, dependency syntactic pattern, dependency syntactic path analysis, and sentence clause analysis etc. These techniques could be implemented in Chinese directly or indirectly. Although Chinese is different from English, for example, Chinese has no space. Many of these languages are the same as each other. For example, part of speech is almost the same. Open domain named entity recognition includes time expression, location, person and organization recognition; sentences are also composed of subject, predicate, object, definition complement, etc. Therefore, transplanting these technologies into Chinese should work well, not to mention Chinese Has made great strides in natural language processing. At present, many Chinese studies have proved its feasibility in [8, 26, 31].
- Disadvantages of English could also be avoided, such as lack of evaluation dataset and criterion. Institutional advantages of our country could mobilize lots of talents to create such kinds of data, so as to provide to lots of researchers as reference.
- Accelerate the development of natural language processing, especially the ability to improve accuracy, speed and domain expansion. From the experience of English open domain entity relation extraction, we can find that relying on syntactic analysis is particularly important for relation extraction, but the premise of this technique is a series of natural language processing techniques. Therefore, natural language processing is particularly important. Considering that relation extraction is mainly applied in the open domain, and speed is particularly important, so speeding up the process of relying on syntactic analysis will help open subject relation extraction.

5 Conclusions

The development of Chinese open domain entity relation extraction research is bound to learn the successful experience of other languages. At the same time, it is necessary to combine the characteristics of this field to develop an effective extraction strategy, and even need a complete set of extraction technology solutions. Entity and relation research will the promote research progress in this area. This paper summarizes and analyzes the extraction of binary and n-ary entity relations in the field of English, and introduces the work related to Chinese, and puts forward some suggestions, hoping to provide reference for domestic research.

Acknowledgement. This research was partly supported by Yunnan Normal University Graduate Research and innovation fund in 2020.

References

1. Anthony F., Stephen S., Oren E.: Identifying relations for open information extraction. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1535–1545. John McIntyre Conference Centre, Edinburgh (2011)

2. Michael, S., Robert, B., Stephen, S., Oren E.: Open language learning for information extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, vol. 4590, Eight Street, Stroudsburgm, United States, pp. 523–534 (2012)
3. Alan A., Alexander, L.: N-ary facts in open information extraction. In: Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction, Montreal, Canada, pp. 52–56 (2012)
4. Luciano, D.C., Rainer, G.: ClausIE: clause-based open information extraction. In: International Conference on World Wide Web, Brazil, pp. 355–366 (2013)
5. Nancy, C., Elaine, M.: MUC-7 information extraction task definition. In: A Seventh Message Understanding Conference, Virginia, pp. 1–53 (1998)
6. Fei, W., Daniel, S.W.: Open information extraction using wikipedia. In: Meeting of the Association for Computational Linguistics, Sweden, pp. 118–127 (2010)
7. Alexander, Y., Michele, B., Matthew, B., Michael, C., Oren, E., Stephen, S.: TextRunner: open information extraction on the web. In: Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, New York, pp. 25–26 (2007)
8. Song, Q., Qi, C.L., Yang, Y.: New Event Relation Extraction Approaches Based on Bootstrapping. J. Commun. Univ. China (Sci. Technol.) **4**, 46–50 (2017)
9. Xiao, L., Daniel, S.W.: Temporal information extraction. In: Twenty-Fourth AAAI Conference on Artificial Intelligence, UK, pp. 1385–1390 (2010)
10. Johannes, H., Fabian, M.S., Klaus, B.: YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In: International Conference Companion on World Wide Web, India, pp. 229–232 (2011)
11. Yao, X.M., Gan, J.H., Xu, J.: Chinese open domain oriented N-ary entity relation extraction. CAAI Trans. Intell. Syst. **14**, 597–604 (2019)
12. Yang, B., Cai, D.F., Yang, H.: Progress in Open Information Extraction. J. Chin. Inf. Process. **28**, 1–11 (2014)
13. Sheng, J.B., Shijia, E., Li, M., Xiang, Y.: Chinese open relation extraction and knowledge base establishment. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **17**, 1–22 (2018)
14. Ndapandula, N., Gerhard, W., Fabian, S.: Discovering and exploring relations on the Web. Proc. VLDB Endow. **5**, 1982–1985 (2012)
15. Sérgio, M., Arrais, J.P., Maia-Rodrigues, J., Oliveira, J.L.: Concept-based query expansion for retrieving gene related publications from MEDLINE. BMC Bioinform. **11**, 212–220 (2010). https://doi.org/10.1186/1471-2105-11-212
16. Roma, Y., Tandan, S.R.: N-ary relation approach for open domain question answering system based on information extraction through world wide web. Int. J. Eng. Appl. Sci. (IJEAS) **2**, 141–144 (2015)
17. Gosse, B., Ismail, F., Jori, M.: Relation extraction for open and closed domain question answering. In: van den Bosch, A., Bouma, G. (eds.) Interactive Multi-modal Question-Answering. Theory and Applications of Natural Language Processing, pp. 171–197. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-17525-1_8
18. Oren, E., Michael, C., Doug, D.: Unsupervised named-entity extraction from the web: an experimental study. Artif. Intell. **165**, 91–134 (2005)
19. Christensen, J., Mausam, Soderland, S., Etzioni, O.: Learning arguments for open information extraction. In: Proceedings of the Sixth International Conference on Knowledge Capture, USA, pp. 1–8 (2011)
20. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th Conference on Computational Linguistics, France, pp. 539–545 (1992)

21. Oren, E., Anthony, F., Janara, C., Stephen, S.: Open information extraction: the second generation. In: International Joint Conference on IJCAI, Spain, pp. 3–10 (2011)
22. Janara, C., Mausam, Stephen, S., Oren, E.: An analysis of open information extraction based on semantic role labeling. In: K-CAP, Ganada, pp. 113–120 (2011)
23. Qin, B., Liu, A.A., Liu, T.: Unsupervised Chinese open entity relation extraction. J. Comput. Res. Dev. **52**, 1029–1035 (2015)
24. Yu, L., Lu, F., Liu, X.L.: A bootstrapping based approach for open geo-entity relation extraction. Acta Geodaetica Cartogr. Sin. **45**, 616–622 (2016)
25. Liu, S.: Chinese entity relation discovery for BigCilin. Harbin Institute of Technology (2016)
26. Li, M.Y., Yang, J.: Open Chinese entity relation extraction method based on dependency parsing. Comput. Eng. **42**, 201–207 (2016)
27. Yang, M.: Research and implementation of Chinese open relation extraction. Nanjing Normal University (2017)
28. Li, Y.: Research and implementation of Chinese open relation extraction. University of Electronic Science and Technology of China (2017)
29. Guo, X.Y.: Entity relation extraction for open domain text. Central China Normal University (2017)
30. Wang, Y., Zhou, G., Nan, Y., Zhen, Z.S., Tian, F.: Open entity relation extraction based on library of relation word. J. Inf. Eng. Univ. **18**, 242–247 (2017)
31. Li, Y.: N-ary Chinese open entity relation extraction. Taiyuan University of Technology (2017)