The Chinese Open Relation Extraction Based on Dependency Parsing

Wang Yuzhao[1, a], Yang Yunfei[1] and Zhao Ruixue[2]

[1]Department of Software Engineering, Jilin University, Changchun,130012,China

[2]School of Computer Science and Technology,Jilin University, Changchun,130012,China

[a]yzwang2014@mails.jlu.edu.cn

Keywords: dependency parsing,chinese relation extraction,heuristic rules,unsupervised

Abstract. Open relation extraction is one of the most promising relation extractioins.However,the precision and recall of the existing methods are far from satisfication especially for extraction from chinese text.In this paper,we propose a method based on dependency parsing, and develope a new system called C-Reverb. Specifically,we design regular expressions for relation phrase through analyzing the Chinese grammar rules.Based on statistics of a large amount of sentences ,we find the distribution between entities and relations to formulate the rules for entity identification. Afterwards, the extracted triples which lack of entities are simply filtered.The experimental results show that our method efficiently improves the precision and recall rate of Chineses relation extraction.Furthermore, relation expressions extracted by our method are more informative than the previous method.

Introduction

With the advent of the big data era ,a vast amount of free text exists on the Web. To help people obtain the information from the massive texts,the task of Relation Extraction (RE) has been proposed,which aims to extract structured information from unstructured or semi-structured texts.

Supervised methods for relation extraction require hunam annotated data , which is laborious and time-consuming.Meanwhile,these methods only target on specific domain,which can not satisfy the need for open domain. To solve this issue, Banko used heuristic rules to process corpus by using Penn Treebank,then trained the TextRunner system to identify the related triples[1]. In 2010, FeiWu[2] matched Wikipedia's infoboxes with texts to generate training data for their model, and improved its precision and recall rate compared with TextRunner.In 2011, Fader[3] introduced two simple syntactic and lexical constraints on binary relations expressed by verbs.Based on this constrain,they developed Reverb system, which performed better than the previous models.

Compared with research works in English, the research on the opening relation extraction in Chinese is insufficient. In 2009, Song Rui employed several types of features to esdablish conditional Condition Radom Field(CRF) model for comparative sentences relation extraction[4]. In 2013, Liu Anan et al., utilized word distance and entity distance constraints to generate candidate relation triples, and then adopted global ranking and domain ranking methods to discover relation words from the relation triple candidate.Finally[5] their model filtered them by using the extracted relation words and some sentence rules.

However, these methods only achieved a low precision and recall rate in multiple sentences of open Chinese relation extraction . In order to solve the above problems, we propose a method based on dependency parsing, and develope a new system called C-Reverb. Specifically,we design regular expressions for relation phrase through analyzing the Chinese grammar rules.Based on statistics of a large amount of sentences ,we find the distribution between entities and relations to formulate the rules for entity identification. The experimental results show that our method efficiently improves the precision and recall rate of Chineses relation extraction.Furthermore, relation expressions extracted by our method are more informative than the previous method.

C-Reverb based on heuristic rules

Identifying relation phrase . We divide the relations into two categories.The first category is the relation as a predicate in the Chinese sentence.Such as "she is our teacher", which "is our class teacher" belongs to this type of relation.The second category is a special structural relation.Such as "red color car", where "red" as the relation between "car" and "color".

Identifying of normal relation phrase.In the Chinese statement, most predicates exist in the form of verb phrases, and the extraction of the first kind of relation is a verb predicate in the sentence.Verb predicate is divided into two categories:

(1) a single verb as predicate

For example: "He left Chengdu", which "leave" as a predicate, describes the relation between "he" and "Chengdu".

(2) verb phrase as predicate

Verb phrases include corrective phrases, complement phrases, interlink phrases, verbal phrases, verbatic phrases, and complex verb phrases. The structure of the verb phrase includes the structure of the action and the structure of the verb. The corresponding lexical relations in sentences are: Verb-Object Relation(VOB),indirect object relation (IOB), complement (CMP), and adverbial (ADV).

We combine with the classification of the two verb predicates to suggest syntactic constraints:

Relation constraint 1: (adverbial +) verb (complement +) (object +);

Where ‘( )’ means it can not appear; ‘+‘ indicates that it can occur for multiple times. The phrase satisfies the condition is also subjected to supplementary constraints.

There are four additional constraints:

Supplemental constraint 1: When the word satisfies the relation constraint 1 and there is a coordinate component (COO), the component should be added into the predicate. For example: "We are singing and dancing in the square." In this sentence,singing and dancing are added into the relation.

Supplemental constraint 2: When a guest object (FOB) appears, it should be added into the relation.

Supplemental constraint 3:When a mediator (POB) structure appears, it should be added into the relation. For example: "She is under the bright lights." In this sentence, "under the ... lights" is added into the relation.

Supplemental constraint 4: When the right attachment (RAD) relation appears, the component is integrated into the corresponding position in the relation.

Supplemental constraint 5: When a fixed (ATT) relation appears, the component is integrated into the corresponding position in the relation.

Identifying of special relation phrase.Through the grammatical structure analysis of a large number of sentences, the following two special relations are summarized.

Relation constraint 2: The sentence "red color car" is in "adjective + noun + noun" structure.The adjective as the relation and the two noun as two entities.

Relation constraint 3: In double structure(DBL),for example, "he called Tom to get his coat." Under the condition of constraint 1, "called Tom to get clothes" as the relation, but "Tom" can be used as the subject of the latter sentence.Only use the relation constraint 1 can not get this relation. Therefore, the introduction of relation constraint 3 is necessary.

Mining related entity pairs . Because the composition of the relation in Chinese is more complicated than that in English,it can not simply follow the ReVerb system to extract the nearest nouns on both sides of the relation as entities.Through the manual labeling method, 1000 sentences were statistically analyzed for the physical location.The results shown in Table 1 are obtained by the relevant information.

Table 1 Distribution of Relation Triples with the Position of Relation Words

Position of Relation Words	Proportion%
Between Entities	71.62%
In the Second Entity	19.43%
Right of the Second Entity	8.95%

Based on this result, we determine the first entity in the sentence by the subject-predicate relation.And then combine the second entity may appear in the relation or on the left side of the relation, we give four entity extraction constraints.

Entity Constraint 1: In the subject-predicate relation (SBV), the majority of the principal components appear on the left side of the relation, and this subject element is used as the first entity .

Entity Constraint 2: In the preposition-object (POB), the phrase as the constituent of the object is used as the candidate for the second entity.

Entity Constraints 3: In the structure of DBL ,the first word after the DBL word is used as the first entity in the next sentence.

Output relation triples. For the existing triples,we do simple filtering to remove the triples lack of entity 1 or entity 2.Then output the remaining relation triples.

Experiments

data and evaluation metrics. We get 500 sentences from the SogouT provided by Sogou Lab and select all relation triples in these sentences by hand,then automatically extract the 500 sentences using the constraints we put forward above.After the experiment,we compare the results with the manual annotation results. The experiment use precision (P) ,recall rate （R) and F to evaluate the experimental results . The formulas of the precision ,recall rate and F are as follows:

$$P = \frac{a}{M} \tag{1}$$

$$R = \frac{a}{N} \tag{2}$$

$$F = \frac{2*P*R}{(P+R)} \tag{3}$$

(where a represents the number of the correct relation triples; M is number of all triples extracted; N represents the number of triples in the standard result set)

Results. After testing the 500 sentences again, get the comparison test results.From the results in Table 3 the improved method in the F value increased by 9.88%, precision and recall rate are significantly improved.These results prove that C-Reverb is further improved,and the extraction effect has been significantly improved.We named the new system as C-Reverb+.

Comparison with the previous method in F value.We compare our method with the method from Qin Bin by experimenting on the same sentences.

Table 2 Comparison test results

Method	P	R	F
Method in this paper	73.40%	87.20%	79.70%
Method from Qin Bin	64.90%	61.20%	62.90%

From the results in Table 2, it can be seen that the F value of our method is 16.8% higher than Qin's , and the precision and recall rate are significantly improved by using the constraints we put forward. This is due to the use of relational instructions in the Qin Bin's method, and the table only contains the relations between people, places, time, which narrow the scope of the relation extraction. However our method can extract not just the relations among people, places and time because we extract the entities according to the position and grammatical structure.

Comparison with the previous method in relational expression.After comparing the completeness of the relationship between the method in this paper and Qin Bin's method,we found the representative sentences in the following table.

Table 3 Comparative example of the relational expression

Method	Sentence	Triples
Method from Qin Bin	Chen Xi director nearly six years for the Jiamusi area to complete some of the first operation	（Chen Xi，operation，Jiamus）
Our method	Chen Xi director nearly six years for the Jiamusi area to complete some of the first operation	（Chen Xi director，some of the first operation，Jiamus）
Method from Qin Bin	He came to Beijing to see a doctor	（He，see a doctor，Beijing）
Our method	He came to Beijing to see a doctor	（He，came to Beijing to see a doctor，Beijing）
Method from Qin Bin	Shen Ping served as dean of the Academy of Sciences	（Shen Ping，served，Academy of Sciences）
Our method	Shen Ping served as dean of the Academy of Sciences	（Shen Ping，served as dean，Academy of Sciences）

As can be seen from the examples in Table 3, the sentence in the use of Qin Bin's method cannot express the integrity of the relation.Qin Bin uses the relation word table for triad extraction, so the table need a high integrity.When the integrity is low, relation expression will be affected.What's more,when there are two relational statements in the correct relation expression,there is no way to judge the relation between the two relational statements because the relation indicated word are in the form of a single word . However the method in this paper bases on the dependency relation between the words to complete the relation expression and rule out the interference factor.The words which modify the center relation words are extended to the relation expression to enhance the completeness of the relation expression.

Conclusions and future work

Based on the method of dependency analysis, this paper proposes method for open Chinese relation extraction. The experimental results show that our method efficiently improves the precision and recall rate of Chineses relation extraction.The heuristic rules will be further refined to improve the precision of extraction.

References

[1] M.Banko,M.J.Cafarella and S.Soderland:*Open information extraction from the Web*.San Francisco:Morgan Kaufmann(2007).

[2] F. Wu and S, D:*Open relation extraction using Wikipedia*. Proceedings of the 48th Annual M eeting of the Association for Computational Linguistics (2010).

[3] A.Fader,S.Soderland and O.Etzioni:*Identifying relation for open information extraction*.Strondsburg,PA:ACL(2011).

[4] Rui Song,Hongfei Lin and Fuyang Chang.Journal of Chinese Information Processing. Vol. 23.No. 2,(2009).(in Chinese)

[5] Anan Liu:R*esearch on chinese open entity relation extraction*.Harbin Institute of Technology(2013).(in Chinese)