DOI: 10.21276/sjet.2016.4.6.1

Scholars Journal of Engineering and Technology (SJET) **ISSN 2321-435X (Online)**
Sch. J. Eng. Tech., 2016; 4(6):268-270 **ISSN 2347-9523 (Print)**

(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Research Article

A Verb-Based Method of Chinese Relation Pattern Extraction

Hu Ruijuan
PLA University of Foreign Languages, Luoyang, 471003, China

***Corresponding author**
Hu Ruijuan
Email: huruijuan01@126.com

Abstract: The current study proposes a verb-based method of relation pattern extraction, which aims to extract entity relations with high accuracy from Chinese website corpus. In this study, entity identifications of human names and island/reef names are carried out by means of ICTCLAS and entity table; accurate sentence examples are generated; and, consequently, verb-based relation patterns are constructed. The experiment results show that the proposed method has good extraction performance.
Keywords: relation extraction; relation pattern; entity identification.

INTRODUCTION

Information extraction is a powerful tool to obtain information, and it is an important means to deal with the serious challenges of information explosion. The goal of information extraction is to extract, from unstructured natural language, structured information which can be understood by computers and of which a main type of structured information is entity relation. Relation extraction is a sub task of information extraction, and the main purpose is to extract entity relations in the sentences [1].

Web has become an information bank containing all kinds of knowledge of mankind; its size is growing at an exponential rate and, in the information it contains, there are varieties of entity relations, such as social relations of people, sovereignty/ownership between states and islands, etc. However, the existing search engine can return only the relevant pages which is contains information that the user concerns, and cannot obtain those varied relation information.

The current study aims at an auto-extraction of inter-entity relations from Chinese websites, and proposes a verb-based method of relation pattern extraction.

RELATION EXTRACTION

Entities in the corpus fall into the following eight categories: names, places, cities, islands/reefs, seas/rivers, organizations, government sections, and military institutions.

Our study focuses on how to extract the relations between these eight categories of entities in order to construct relation patterns. In this paper, we conduct studies on the extraction of relations between names and islands/reefs.

The relation extraction strategy as is proposed includes getting sentences and constructing relation patterns. The principle of entity relation extraction is shown in figure 1.

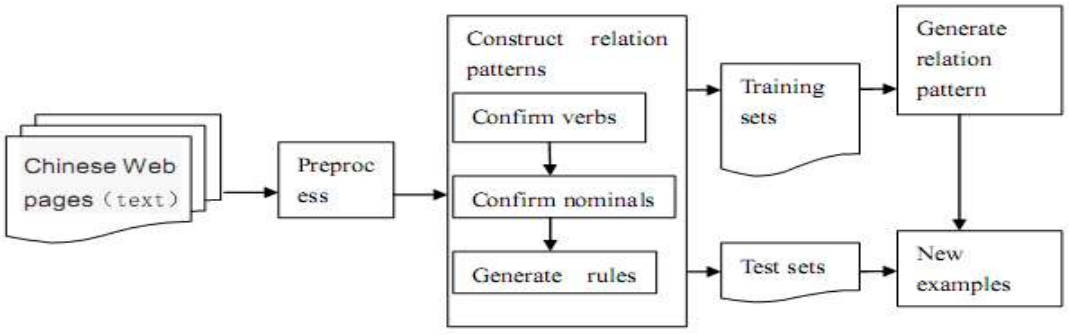

Fig-1: The principle of entity relation extraction

Sentence acquisition

The texts of Chinese Web pages are pre-processed so we get those sentences which contain the entities and which serve as sentence examples for relation extraction. The pre-processing includes paragraph division, sentence segmentation, POS tagging, and named entity recognition [2]. Since the current study deals with the relations between figures and islands/reefs, we use ICTCLAS to identify these two types of entities in the segmented sentences, where figures are tagged '[1-figure xx]' and islands/reefs are tagged '[23-island xx]'.

The construction of relation patterns

Traditional heuristic methods tag POS of words in the sentences, and construct patterns by replacing positions of entity pairs with wildcards [4]. These methods, however, lack compatibility and accuracy. For example, the result of pattern construction of "北京是中国的首都 (Beijing is the capital of China)" is " object是/v target的/u首都/n ", whereas in the pattern construction of "北京是中国政治文化的中心 (Beijing is China's political and cultural center)", there is no demonstration of the relationship between Beijing and China: capital-of.

The method of relation pattern extraction as is proposed by this paper studies the relation patterns in the context before and after the entities in the seed sentences. The process of studying is not that of mechanically treating all the words that appear before, in and after the entities as featuring words of the studied patterns [5]. Studies of the sentences show that there are a lot of modifying words (adjectives, adverbs, exclamations, among others) in the context connecting two entities. If all the words are treated as featuring words of the extracted patterns, then, on the one hand, the length of the pattern is greatly increased, which takes too much machine time in the pattern matching that follows; on the other hand, the accuracy of long pattern matching will be undermined. Therefore, this paper puts forward a verb-based method of relation extraction, which takes "verb" as the core, and puts together sentences with the same verb and extracts their relation patterns.

The following is an example of sentence tagging:

动词：到达

从谭门镇出发，将航向调到东南110度，3天3夜后"琼琼海08068"号渔船船长[1-许卫]可到达[23-黄岩岛]。

The relation being extracted is that between (concepts of) entities. The contextual realization of this relation is context constraint. We summarize various context constraint rules and apply them within language use. The planned calculation breaks of contexts include: SENT, meaning in the same sentence; ORD, indicating a sequential order; DIST_n, referring to a distance no more than n, etc.

Below is an example of relation pattern:

CONCEPT: ACTION_ARRIVE: 到达

CONCEPT: NAME: 许卫

CONCEPT: NAME_ISLAND: 黄岩岛

MCONCEPT_RULE:ARRIVE(person,island):(SENT,ORD("_person{NAME}", "ACTION_ARRIVE", "_island{NAME_ISLAND }"))

The relation pattern consists of two parts: concept and rule. "CONCEPT" represents the concept and includes verbal concepts and nominal concepts. Verbal concepts start with "ACTION_", and nominal concepts refers to entities, such as people, places (countries, islands, seas, cities), organizations, etc.; "CONCEPT: NAME: 许卫" refers to the name "许卫"; "CONCEPT: NAME_ISLAND: 黄岩岛" refers to the name of the island "黄岩岛". Extraction is aimed at the relations between the concepts. We define extraction rules with "MCONCEPT_RULE", the rule being named with a verb "ARRIVE" and composed of two parameters, person and island, which correspond to nominal concepts of NAME and NAME_ISLAND.

EXPERIMENT RESULTS AND ANALYSIS

Experiment data

The subject of experiment is the relation between human names and island/reef names. A total of 21,467 Chinese Web pages are used, whose paragraphs are divided and whose sentences are segmented. We got 435,210 different sentences. With these sentences we used ICTCLAS to carry out entity identification and entity table comparison, and finally collected 7,959 sentences containing names of people and of islands, which constituted the experiment set. All the sentences were grouped according to the "verb" they contained. Named entities (person names, island and reef names) were tagged. We selected 100 representative sentences as the seed set relation patterns, with the remaining sentences being the testing set.

Experiment results

The process of constructing the relation patterns is realized by the relation extraction system, as shown in Figure 1. We input the verb "indicate" to look for related sentences, defined the rule name "IMPLY", generate the verbal concept "ACTION_IMPLY", extract nouns from the sentences and generated the nominal concept, thus constructing the relations between concepts and get the relation patterns (i.e., rules).

The constructed relation patterns were written into Txt files.

Since the experiment studies a particular type of relation, we use two criteria: accuracy and recall rate [6]. The formulas are as follows:

Accuracy P= (the number of a particular entity relation accurately extracted) / (the number of a particular entity relation extracted in a testing set)

Recall rate R= (the number of a particular entity relation accurately extracted) / (the number of a particular entity relation that the testing set is supposed to have)

The results of the experiment are shown in Table 1.

Table 1: Experiment results of relation pattern extraction

DIST	Extracted person-island relations	Correct relations	Due person-island relations	Accuracy	Recall rate
DIST_2	132	105	169	79.5%	62.1%
DIST_4	98	75	114	76.5%	65.8%
unlimited	146	101	143	69.2%	70.6%

Table 1 suggests that the recall rate of the verb-based method of relation pattern extraction, as well as the accuracy of the extracted person-island relation, is closely related to the DIST value. DIST_2 refers to a distance no more than 2, that is, when the distance between 'person' and the verb is less than 2, the accuracy of the extracted relation reaches up to a relatively high percentage of 79.5%, while its recall rate is comparatively low; when DIST is bears a value of DIST_4, the accuracy of the extraction is 76.4%, lower than that of DIST_2, and its recall rate is higher than DIST_2; when the value limit of DIST is removes, in other words, when the distance between 'person' and the verb can be as far as possible, the extraction accuracy is much lower and its recall rate is rather high.

CONCLUSIONS

Ordinary pattern matching methods extract entity relations by matching patterns with sentences and the words in the sentences in terms of format or form. Such methods generally demand regularity of the syntax and structure of the sentences. However, for Chinese Web corpus which contain various relations and have relatively flexible text structures, ordinary methods of pattern matching do not perform as efficiently as expected. Therefore, this paper puts forward a verb-based method of pattern matching to extract the relations between entities in sentences, fulfilling the task of extracting the relations between human names and island/reef names. Future work may be focused on further mining of the relation instances in order to increase extraction accuracy and obtain more semantic information; future research may also extract relations of other types.

REFERENCES

1. Chinese Academy of Institute of Computing Technology. ICTCLAS Chinese Analysis System; available from http://ictclas.org/.
2. Medelyan O, Milne D, legg C; Mining Meaning from Wikipedia. International Journal of Human-Computer Studies, 2009; 67(9): 716-754.
3. Agichten E, Gravano L; Snowball. Extracting relations from large plain-text collections. Proceedings of the fifth ACM conference on Digital libraries, New York, ACM Press, 2000; 85-94
4. Zelenko D, Aone C, Rechardella A; Kernel Methods for Relation Extraction. Journal of Machine Learning Research. 2003; 3:1083-1106.
5. Sundaresan N, Yi J; Mining the Web for Relations. Computer Networks. 2000; 33(1-6):699-711.