Journal of Theoretical and Applied Information Technology
31st December 2012. Vol. 46 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN IMPROVED SELF-LEARNING MODEL BASED SOCIAL RELATIONSHIP EXTRACTION

CHONGWEN WANG, TONG SHEN, YI HUANG

School of Software, Beijing Institute of Technology, 100081, Beijing, China

ABSTRACT

How to extract social relations from the text content in internet is a problem. A supervised method based on machine learning algorithm has been used to solve the problem. Based on the characteristics of social relationship, the appropriate rules have been made for feature extraction. Based on the result of feature extraction, two methods have been proposed which are support vector machine (SVM) and the maximum entropy model for the relation extraction experiment. The results show that support vector machine algorithm is better than the maximum entropy model.

Keywords: *Social Relation, Relation Extraction, Support Vector Machine, Maxent Model, Machine Learning*

1. INTRODUCTION

With the rapid development and large-scale popularization of the Internet technology, the data of information society showed explosive growth. Search engine technologies developed fast and rapidly in this background. Google, Yahoo, Baidu, Bing and other search engines appeared and quickly occupied the market by the ability to meet the requirements of locating the required data in the vast information ocean. But in the other view, the current search engine still only processed and analyzed simple data from Internet or calculated and ranked the page structure and links. It is still an empty field to in-depth understand or mine data and extract knowledge behind the large-scale data, especially for social relations extraction and discovery. More and more researchers, university labs and technology companies focused in the study of social relation extraction.

Relation extraction technology is a key point in the information extraction research that organized the natural language text into a structured or semi-structured data. The main task of information extraction is to extract and store the formatted information from the original data source for the subsequent retrieval and processing. The research of relation extraction that is based on information extraction takes up the relationship between physical objects. The relationship modeling, relation defining, corpus capturing and extraction algorithm are the hot topic in current relation extraction research field. With the Internet technology quickly developed and deeply influence on human society, the model and network of human social relations are gradually transferred to the virtual transition area of Internet. We can organize and retrieve large-scale data and information on the internet through the social relation network. And another intuitive instance is the popular research of social network that is in a key position with relation extraction. Now people can announce message through SNS and microblogging conveniently. Based on this information it is important for theory research and practical application to map out the social relation network.

In this background, we focused on the research of the relevant algorithms and solutions for social relation extraction especially on the supervised and unsupervised machine learning method, relational feature vector extraction, and relational model training and so on. The research aims to improve social the efficiency and performance of the algorithm for social relation extraction.

2. CORPUS CONSTRUCTIONS

2.1 Social Relation Types

The ACE evaluation conference defined seven major relation types such as institutional relation, part and whole relation, human and social relation and so on. In each relation category a number of subcategories are defined. The hierarchical definition of the entity relation type is same with the nature and society relational model. In this paper, the social relation extraction is one part of the entity relation system. The ACE08 evaluation conference gave 510 records of the Chinese social relation. The database is too small to meet the requirement of social relation extraction task. And

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the ACE only defined three simple subtypes for the social relation type that was business, family and lasting-personal. Based on the definition, we subdivided the social relation type and then collected and labeled our own Chinese social relation corpus.

The ACE08 evaluation conference only defined three subtypes for the social relation category. In this paper we refer to the relation types that defined by ACE conference then redefined the social relationship categories. There are six social relation types that are relatives, friends, lovers, co-operation, subordinates and idols.

2.2 Corpus Processing

2.2.1 Corpus source

Now many researches of natural language processing chose the Internet as the source of corpus. Since this relationship involved extraction experiments do not involve real-time data, so this choice of a laboratory to provide Internet search dog text classification corpus, we further search through the database of dogs available all documents, read the contents and of storage into a tree structure in XML format, so that we can in the next phase of more in-depth text corpus lexical processing.

2.2.2 Corpus processing

The original format of corpus was webpage document on the Internet. The document needs in-depth lexical processing operations, such as word segmentation and name recognition, and so on. Meanwhile we will take a number of rules and methods to filter the document content. The Filtering rules would be described as below.

Firstly, we filtered out the sentence that was shorter than the required length. According to the results of the segmentation, we could get all of the words in the sentence. We defined the length as the count of the words. The filtered threshold was defined as 10.

Secondly, we filtered out sentences that did not contain two names. Apparently social relation extraction needs at least one person involved. In this paper the character entity appeared in a sentence that reduced the problem complexity.

After taken two filtered strategies, we can get a group of sentences that could be used for relation extraction. The second filtered strategy needs the accurate recognition for the character entity that involves the related research of Chinese name recognition and identification. In one word, the corpus processing includes a series of work such as sentence selection, word segmentation, and POS tagging, Chinese name recognition and so on.

2.2.3 Corpus annotation

The original text was processed with the preliminary lexical operation and name recognition. And then we filtered the original data by two rules and got a number of required sentences. These sentences contained at least two character entities and human could understand the sentence. Based on such preconditions, we defined the following rules for the relation annotation.

Firstly the relation annotation accepted the sentence as annotation atomic unit. If there were several character entities in one sentence, we took the first two entities which were defined as Person1 and Person2. If the Person1 and Person2 referred to the same character entity, the corresponding sentence was filtered out.

Secondly the sentence needs to contain the social relation which reflected in the sentence directly. The social relation could be described by the special feature word directly or be reflected by the semantic context indirectly. For some well-known social relationship, it was not marked if it did not reflect in the sentence.

2.2.4 Format of corpus

In the corpus acquisition process the temporary data need to be stored and the annotated results need to be stored. The original text was grabbed from Internet and then the paragraph was processed with word segmentation and POS tagging and the generated results were in line with the tree structure. So in this paper we proposed to use XML format to store data.

In this paper, we referred to the document format of relation extraction task in ACE evaluation conference.[2].And according to the characteristics of the relation corpus, we improved the document format. For example, we added parentId which referred to the sentence ID of the source text. The relation type referred to the six relation types from the number one to six. Person1 and Person2 nodes referred to the two character entities in the sentence.

3. THE SUPERVISED METHOD FOR RELATION EXTRACTION

3.1 The Establishment Of The Relation Model

3.1.1 Characteristics of relation system

In this paper we defined six social relation types. Each relation type could regard as a category, and the relation extraction task could be transformed into a complex pattern classification problem. For instance, the entity pair (Person1, Person2) could be

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

classified into one relation type. Thus we need focus to the characteristics of relation types and determine which characteristics take a greater role in classification.

Firstly the relation expression of the entity pair(Person1,Person2) was usually depended on a certain type of feature words. And these feature words could appear between Person1 and Person2, or before Person1 or after Person2 as the description of character. These feature word is usually to be noun or adjective as the description of character entity.

Secondly the position of feature description may be far apart from Person1 and Person2. The speech is usually a verb and the social relation can be derived from a particular action or behavior. For example, the words marriage and marry could describe the couple's relationship as the characteristics word. Its position in the sentence is often apart from Person1 and Person2. So such features may be ignored if only focus the words around for Person1 and Person2. The features need to be paid attention in the feature extraction process.

Thirdly the sentence dependency syntactic between character entities needs to be considered. The feature word could reflect the relation type but could not determine whether a relationship exists between character entities. By the dependency syntax features, we can describe the relation between the character entities and help to extract the social relations.

3.1.2 Feature extraction method

The establishment of social relations model needs to extract the feature vector from the original text corpus. In this paper we extracted the attributes which associated with the character entity and organized as feature vectors. Meanwhile the character entities were defined as Person1 and Person2 and the position of Person1 was before Person2. All the eigenvalues would be combined into a feature vector as the format of <w1,w2...w31> and extracted feature from processed corpus.

3.2 Analysis of algorithms and experimental results

3.2.1 Experimental Data

In this paper we use sogou corpus as our experimental data. The data source is from sohu news. We selected 2/3 of data as training set randomly and the remaining as the test set. There were 1,350 sentences as the experimental data to test classification results.

In our annotation data, we not only marked the entities and the entity attributes, but also marked the entity relations and the relation attributes. The data and annotation results were stored by XML format. In the experimental data, each two entities in a sentence form a relationship instance.

3.2.2 Evaluation Standards

In the ACE evaluation conference, there are two indicators to evaluate the social relation extracted results which named recall and accuracy[5], The value of recall and accuracy would be different in some case. So we need to compute the weighted average value of the recall and accuracy and the resulted value was called F-value. The formula was defined in Equation 1. The weighted value represented the important level for the recall and accuracy in the F-value computing process. By the calculated results of recall and accuracy we could quantify and analyze the experimental result of social relation extraction result.

$$F - Score = \frac{(w^2 + 1.0) * \Pr ecision * \Re call}{(w^2 * \Pr ecision) + \Re call} \quad (1)$$

3.2.3 Experimental results and analysis

The maximum entropy model used the development package of Maxent 3.3.3k which is provided by Princeton University research laboratory and other joint company. And the source code and documentation is provided on its website. We took experiment in the test sets of relation corpus. And then we got the classification results of social relation types. At last we calculated the recall value, accuracy value and F-value. The experimental results were shown as the following Table 1.

Table 1: The Experimental Results Of Maximum Entropy Model

Categories	Precision (%)	Recall (%)	F-Score (%)
Relatives	74.50	74.30	74.40
Friends	85.60	82.10	84.10
Lovers	73.40	75.50	74.50
Partnership	61.40	62.20	61.90
Subordinates	73.20	69.50	71.10
Idol relationship	76.10	73.20	74.60

In this paper we took the Libsvm development kit that provided by Taiwan University professor Lin for the support vector machine algorithm. The Libsvm development kit provided a framework of variety environments and languages oriented. For example the libsvm could support the Matlab platform, the python language, the java language

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and so on. The input of Libsvm needs to be numbers. So we used a hash table to map the entire string variable to the index number. Then the string feature vector was transferred to the number. Eventually the experimental results were shown in Table 2.

Table 2: Experimental Results Of Support Vector Machines

Categories	Precision (%)	Recall (%)	F-Score (%)
Relatives	78.30	86.11	82.10
Friends	83.80	90.76	87.50
Lovers	69.40	83.52	76.00
Partnership	67.20	65.12	64.60
Subordinates	77.50	68.72	72.70
Idol relationship	76.40	85.44	80.60

According to the experiment results the support vector machine algorithm was better than the maximum entropy model algorithm on the recall value or the accuracy value. But there was not material difference between the two algorithms. And the support vector machine algorithm was more complex and difficult than the maximum entropy model algorithm. The maximum entropy model algorithms support the string feature vector. So in some case we also would prefer to the maximum entropy model algorithm.

The key point of supervised learning method was hot to build an effective model. So we analyzed the characteristics of social relations firstly and then proposed how to extract features in the sentence effectively. The extracted features not only included the description of character entities but also considered the effect of some described verb. Based on the feature extraction of corpus, we took experiment with support vector machine algorithm and the maximum entropy model algorithm. Then we compared the experiment results and concluded that the effect of support vector machine algorithm was better than the maximum entropy model algorithm in the same experiment conditions.

4. THE UNSUPERVISED METHOD FOR RELATION EXTRACTION

4.1 The algorithm workflow

Then relation triple <e1, e2, R> was the system input. There were two known items in the triple such as entity e1 and entity e2 for example. Then we can transfer the relation extraction problem to an answer extraction problem. With the help of search engines and knowledge-based question answering system, we could complete the relationship between the triples. The algorithm workflow included query construction, query expansion and answer extraction.

4.2 The Algorithm Design

4.2.1 Query construction

We selected the entity e1 and the corresponding characteristics of the relation fw to combine an identified tuple <e1,?,fw>.For example, <Jiang Jieshi, ?, Son>, <Jiang Jieshi, ?, child>, <Song Meilin, ?, marry> and so on. And according to the different speech of the characteristics word of the relations, we need to construct the query questions by different heuristic rules.

4.2.2 Query Expansion

According to the different speech of feature words, we constructed the different query questions. Based on the query questions and with the help of answering system, we expanded the content and the scope of the inquiry questions.

We searched the constructed questions in the answering systems. Then we could get the sorted results according to the relevance of the query questions. And in the answering page, there were five recommended issues that related to the query questions. We took the related questions as the expanded questions and at last there were twenty-five query questions in total.

4.2.3 Answer Extraction

After the query construct and query expansion, we proposed a frequency-based statistical method to extract answer. The algorithm took the relational tuple as unit and according to the entity type we got the statistics for the entity types in the text that was named the frequency of the candidate answers. The credibility of candidate answer that was corresponding to the entity e2 was shown in Equation 2.

$$Conf_{fg}(e2_i) = Freq(e2_i) \quad (2)$$

The entity e2 selected the candidate answer which was the maximum credibility and greater than the threshold minFreq.

4.3 The Experiment Results And Analysis

4.3.1 The experiment data

The experiment was based on the entity pair "name - name". The seed entity and the characteristics of relation type were obtained from the social relation type system and the labeled corpus. We selected 500 seed entities and six relationships to make the experiment that included

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

relatives, friends, lovers, co-operation, subordinates and idols.

4.3.2 Experimental results and analysis

We chose the frequency statistics method as the answer extraction rule and tested the labeled corpus. The experiment result was shown in Table 3.

Table 3: Experimental Results Of Unsupervised Methods

Categories	Accuracy rate(%)
Relatives	52
Friends	62.
Lovers	58.
Partnership	46.
Subordinates	77.
Idol relationship	75.
Average	61.5

The web data mining method for relation extraction could achieve a good result in the small-scale experiment. Compared with the supervised learning algorithm, the accuracy of the seed extraction method was lower than the support vector machine algorithm or the maximum entropy model algorithm. But this method did not require the pre-labeled corpus and could be constructed their own queries and extracted the answers. So there was still a sense of practical value for the algorithm.

5. CONCLUSION

In this paper we focused the research of the social relation extraction in the Chinese language field. Entity relationship extraction was the subtask of the information extraction and was one of the important unresolved problems in natural language processing field. Its main task was to extract the semantic relations between two or more entities. There were several researches in the common domain field of entity relation extraction and there were a wealth of corpus resources existed. But in the social relation field, there was lack of existed research and corpus. So in this paper we started from the construction of the corpus to take a systematic research on the social relation extraction.

Firstly we built a Chinese social relation corpus. The first step of building a corpus was to determine the social relation types. In this paper we defined six social relation types and proposed a detailed specification of the corpus annotation. In order to ensure the quality of the corpus, we took the ACE corpus building process as reference.

Secondly in this paper we proposed machine-learning algorithm for entity relationship extraction. The entity relation extraction was regarded as a classification problem. And we used support vector machine and maximum entropy model algorithm for training and testing. The experiment results showed that the support vector machine algorithm was better than the maximum entropy model algorithm in the same experimental condition. In the six social relation types, the support vector machine algorithm obtained the average accuracy of 76.40%, the average recall of 85.40% and the average F-value of 80.60% that was reached the practical level. Since the accuracy of the machine learning method was low, we need to improve the feature extraction or enhance the kernel function in the future research.

Finally for the lack of labeled corpus extraction problem, we proposed a relation seed extraction method that based on the web data mining. The algorithm transfers the relation extraction task to the factual answer extraction problem. The basic query was constructed by the simple heuristic rules. With the help of the answering system, we expanded the basic queries. Then we retrieved a lot of page abstract documents from the answering system. At last we use the frequency statistics method to extract the answer and fill the relation. Finally, we made experiment in the six relation types and achieved accuracy value of 61.5.

REFERENCES:

[1] Sun. A, R. Grishman, S. Sekine, "Semi-supervised Relation Extraction with Large-scale Word Clustering", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics(ACL 2011)*, 2011, pp. 521-529.

[2] Yee Seng Chan, Dan Roth. "Exploiting background knowledge for relation extraction", *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, Association for Computational Linguistics, 2010, pp. 152-160.

[3] Banko. M, Cafarella. M. J, Soderland. S, Broadhead. M, Etzioni. O, "Open information extraction from the Web", *Proceedings of the International Joint Conference on Artificial Intelligence*, 2007, pp. 2670–2676.

[4] Banko. M., Etzioni. O, "The tradeoffs between open and traditional relation extraction", *Proceedings of the Annual Meeting of the ACL*, 2008, pp: 28–36.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[5] Zhu. J, Nie. Z, Liu. X, Zhang. B and Wen. J-R, "Statsnowball: a statistical approach to extracting entity relationships", *Proceedings of the International Conference on World Wide Web*, 2009, pp. 101–110.

[6] Ang Sun, "A Two-stage Bootstrapping Algorithm for Relation Extraction", *Student Research Workshop*, RANLP 2009, 2009, pp. 76-82.

[7] Yuhang Yang, Qin Lu, Tiejun Zhao, "A Clustering Based Approach for Domain Relevant Relation Extraction", *International Journal of Information*, Vol. 12, No.2, 2009, pp. 399-410.

[8] F. Mesquita, Y. Merhav, D. Barbosa, "Extracting information networks from the blogosphere: State-of-the-art and challenges", *Proceedings of the 4th Int'l AAAI Conference on Weblogs and Social Media*, 2010.

[9] Banko. M, Cafarella. M. J, Soderland. S, Broadhead. M, Etzioni. O, "Open information extraction from the Web", *Proceedings of the International Joint Conference on Artificial Intelligence*, 2007, pp. 2670–2676.

[10] Chapelle O, "Training a support vector machine in the primal", *Neural Computation*, Vol. 19, No. 5, 2007, pp. 1155-1178.