Journal of Computers Vol. 30 No. 1, 2019, pp. 31-41
doi:10.3966/199115992019023001004

Research of Unsupervised Entity Relation Extraction

Yun Liu[1*], Mingxin Li[1], Hui Liu[2], Junjun Cheng[2], Yanping Fu[3]

[1] Department of Electronic and Information Engineering, Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, China
{liuyun, 15125021}@bjtu.edu.cn

[2] China Information Technology Security Evaluation Center, Beijing, China
{liuhui, chengjj}@itsec.gov.cn

[3] Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, China
17111012@bjtu.edu.cn

Received 9 January 2018; Revised 22 January 2018; Accepted 22 January 2018

Abstract. Nowadays, the Internet is growing rapidly and the number of network data is also growing rapidly, which makes it more difficult to get information from the massive network data with traditional methods. Entity relation extraction is an important research direction of natural language processing. It can find and identify the semantic relation of the entity to analyze the abstract textual data. Unstructured network data can also be transformed into structured data by using Entity relation extraction. This paper presents an unsupervised entity relation extraction model, which can overcome the shortcomings of traditional methods, such as needing a lot of man-made work and poor portability. In this model, we create a filter function at first. Then we extract the relational feature words by using context window and parsing. We use affinity propagation clustering algorithm to get the relation of entities, which can obtain better results than k-means clustering algorithm.

Keywords: conditional random field, entity relation extraction, information extraction, named entity recognition, parsing

1 Introduction

With the rapid development of network and computer technology in recent decades, massive network information has brought great challenges to the organization, search and analysis of information. Turning the unstructured text to the structured text can help understand text content by annotating the semantic information. The task of the entity relationship extraction has been extensively studied in different methods [1-3]. There are three main ways to solve the problem of entity relationship extraction including the method of based on constructing rules [4], the method of using of machine learning [5-7] and the method of based on the feature [8-10].

The method of constructing the rules requires the constructor to have a deep understanding of the characteristics of a certain field and applies the rules or templates summarized by manual or machine learning. Then it uses the template matching method to extract the entity relationship. However, with the manual or machine annotation, the cost is large and the transplantability is poor, so it is gradually replaced by the method of machine learning [11]. In the method of machine learning, the current mainstream method is based on a supervised approach, which requires training data and complex feature extraction techniques and prior definition of relationship type systems, often requiring a large number of manual annotation work [12-13]. The method of the feature extracting is simple and effective, its main idea is to extract useful information including lexical information and grammatical information from the

* Corresponding Author

context of the relationship sentence instance as a feature. The eigenvector is constructed by calculating the similarity of eigenvectors to train the physical relationship extraction model. The key of this method is to find the distinguishing feature of the class, constitute the multi-dimensional weighted feature vector, and use the appropriate classifier for classification.

According to the degree of dependence on labeled data, the method of physical relationship extraction can be divided into supervised learning method, semi-supervised learning method, unsupervised learning method and open extraction method [14]. Supervised learning method is the most basic entity relation extraction method, its main idea is to train the machine learning model on the basis of labeled training data, and identify the type of the relationship between the test data [15]. The unsupervised entity relationship extraction method does not need to rely on the entity relationship annotation corpus, and its realization includes two processes of relation instance clustering and relationship type word selection. Firstly, the entities with some degrees of similarity are grouped into a class according to the context of the entity, and select the representative words to mark the relationship. Literature [16] firstly proposed the method of unsupervised entity relationship extraction in the association for computational linguistics (ACL) meeting, laying the foundation for the unsupervised entity relationship extraction. Literature [17] firstly proposed a semi-supervised entity relation extraction method based on Bootstrapping, this method summed up the sequence model of entity relationship from the context containing relationship seeds, then using sequential patterns to find more relations seed instance and form a new relationship between seed collection. This method has high accuracy of sequence pattern, but the recall rate is relatively low. In order to improve the recall rate of sequential pattern, the scholars introduced the concept of soft-pattern, and generalized the elements of the composition pattern, which improved the recall rate of the sequential pattern to a certain extent. The basic hypothesis of the open entity relation extraction method: if the known two entities have a semantic relation, all sentences containing the two entities potentially expressed the semantic relationships between them [18]. The Open entity relation extraction maps quality entity relationship instance to large-scale text by entities knowledge base (such as DBPedia, YAGO, OpenCyc, FreeBase or other domain knowledge base), obtains the training data according to the text alignment method and use the supervised learning method to solve the problem of relation extraction. However, the training corpus obtained by this method has more noise.

In this paper, we use the unsupervised entity relation extraction algorithm to extract the entity relation based on the use of the characteristic words, which can solve the problems of man-made annotation and poor portability. This method does not need the data training, which can solve the big drawback of human input, and the use of content covering a wide range of news data, effectively solve the problem of domain adaptability. This method does not need to determine the type of relationship in advance, can solve the problem of definition of relationship type [19].

We have also made some improvements. This paper presents a screening method for the entity relationship pair, and also proposes a method based on “window” and syntactic analysis to extract feature words. This paper improves the performance of the method through these improvements. We evaluate each method experimentally, demonstrate their synergy, and compare our unsupervised entity relationship extraction model with artificial annotated feature words model. Our main contributions are:

(1) This paper demonstrates its feasibility of implementing unsupervised, domain-independent information extraction from the entity relationship with high precision. Much of the previous work on information extraction focused on small document collections and required manual annotation examples.

(2) This paper presents the comprehensive overview of the improved unsupervised entity relationship extraction model. We describe the process of building model including pre-treatment, filter, extraction and cluster.

(3) This paper show that excellent clustering results because of the effective feature words with combining the “window” method and syntax analysis method.

(4) This paper describe and evaluate the superiority of affinity propagation (AP) clustering algorithm in terms of entity relationship extraction.

2 Improved Unsupervised Entity Relationship Extraction Model

In unsupervised relation extraction, the information of the expected results cannot be obtained in advance, the method of marking entity relationship can express the characteristics of the clustering results, so how to present the relationship is the main content of research. The relationship between named entity pairs is

expressed through the context between them and the relationship can be obtained through the analysis of the context. The basic idea of unsupervised machine learning is based on the theory of distributed assumptions, that entities with same relationship will have similar contextual content, and a representative vocabulary can be used to describe the relationship.

In this paper, we need to carry out the following steps when extracting entity relationships. Firstly, we describe the pre-treatment stage. The pre-processing is used to segment the text, recognize named entities, and do syntactic analysis and annotation [20-21]. Secondly, in order to determine the degree of entity correlation and determine the named entity with entity relationship, we use entity relationship filtering. Thirdly, for entity pairs with entity relationships, we select the feature words which can represent the entities' relationship through relational feature words extraction. At last we use the AP clustering algorithm to cluster the feature words with word vector construction and words clustering. Then by the processes of pre-treatment, filtering, extracting and clustering, we can get the relation of the entities.

The flowchart of unsupervised entity relation extraction is as Fig. 1.

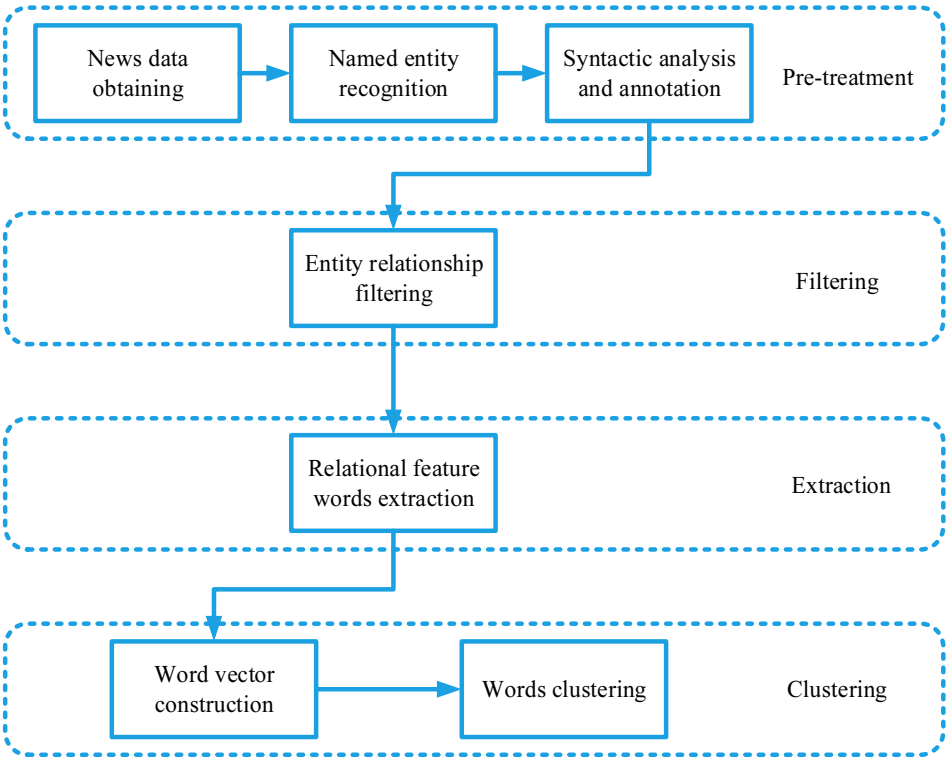

Fig. 1. Flowchart of unsupervised entity relation extraction

2.1 Named Entity Recognition

The Named entity recognition is the basis of entity relationship extraction. In this paper, we use a named entity recognition method based on conditional random field (CRF). CRF is a conditional probability distribution model of the output random variables under the condition of given a set of input random variables, its characteristic is to assume that the output random variables constitute a markov random field. It has the same characteristic index weighting as the maximum entropy model, but the training process uses a complete, non-greedy search algorithm, which is very effective. The method sets different recognition templates depending on the entity types, and it uses calibration rules for geographic entities and organizational entities [22-23].

Improved architecture of named entity recognition is as Fig. 2.

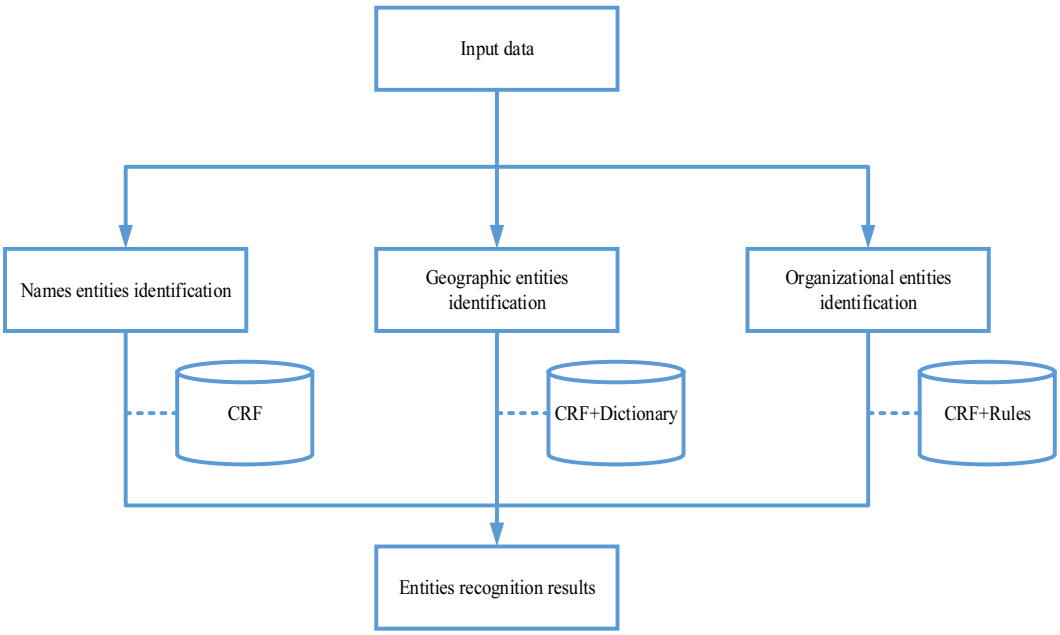

Fig. 2. Improved Architecture of Named Entity Recognition

In our improved architecture of named entity recognition, input data can be managed by three different types with names entities identification, geographic entities identification and organizational entities identification. These identification types respectively use CRF algorithm, CRF and dictionary algorithm and CRF and rules algorithm to obtain final entities recognition results.

The advantage of the above architecture is that for different types of named entities, you can use specialized algorithms and specialized external dictionaries to identify, which can make the recognition to achieve higher accuracy.

In this paper, the recognition process for geographic entities is based on the use of CRF algorithm, using the dictionary as a calibration rule for place name entity recognition. This part of the data mainly come from the "World Translator Dictionary" and the Chinese name dictionary from the Internet.

For the identification of organizational entities, the identification of composite organizational entities is relatively more difficult. Therefore, this paper uses the CRF algorithm to combine the rules to identify the organizational entity [24].

The specific rules are as follows:

- When a subject or object in a sentence is composed of multiple types of words and contains the suffix of the organization entity, the subject or object can be directly determined as a composite organization entity.
- When the syntax is parallel to each other in two parts, one of which is a simple organizational entity, the other part of a number of types of words, and which contains the organizational body of the suffix words, this part can be directly identified as a composite organization entity.

Through the use of the above two rules, the composite organization entity can be effectively screened. Combined different rules with the CRF algorithm to determine the subject or object, we can get the organization entity recognition results.

2.2 Entity Relationship Pairs Filtering

Naming entity relational extraction needs to identify a named entity pair that may have a relationship first. In this paper, we use the title to dig a pair of named entities that may have a relationship. After the named entity recognition, it is possible to get all the words in the data that are identified as named entities. After the data is filtered, the named entities are combined to get the pending entity pairs.

This paper presents a standard for entity relevance evaluation to measure the existence of a relationship between two named entities:

$$S(w_1,w_2)=\frac{x(w_1,w_2)\left(x(\mathrm{w}_1)+x(\mathrm{w}_2)\right)\left(x^2(\mathrm{w}_1)+x^2(\mathrm{w}_2)\right)}{x^2(w_1)x^2(w_2)} \quad \textbf{(1)}$$

In the above formula, $S(w_1,w_2)$ represents the degree of association between the named entity w_1 and w_2; $x(w_1,w_2)$ represents the number of news text sources that the named entity w_1 and w_2 appear together; $x(w_1)$ and $x(w_2)$ represent the number of text sources that have named entities w_1 and w_2.

Then, the threshold $S_\tau(w_1)$ of the named entity w_1 is calculated from the training data and the processed data:

$$S_\tau(w_1)=\min\left\{S(w_1,w_{11}),S(w_1,w_{12}),\ldots,S(w_1,w_{1n}),\ldots\right\} \quad \textbf{(2)}$$

In the above formula, $w_1,w_{12},\cdots$ are the named entities that have determined the relationship with the named entity w_1.

At last, according to $S_\tau(w_1)$ and $S_\tau(w_2)$, which are the thresholds of the named entities w_1 and w_2, we can determine whether the entity pair satisfies the relation extraction condition:

$$f(w_1,w_2)=\begin{cases}1, & S(w_1,w_2)\geq\dfrac{S_\tau(w_1)+S_\tau(w_1)}{2}\\ 0, & \text{otherwise}\end{cases} \quad \textbf{(3)}$$

In the above formula, when $f(w_1,w_2)$ is 1, it is concluded that there is an entity relationship between named entities w_1 and w_2; When $f(w_1,w_2)$ is 0, it is concluded that there is no entity relationship between named entities w_1 and w_2.

In this way, we filter out the named entity pairs with low degree of association to reduce the workload of subsequent processing.

2.3 Relational Feature Words Extraction

The key of the Named entity recognition is to extract their relevant features and define entity feature by analyzing the characteristics of named entities. At present, in the extraction of entity relations, the methods of relational feature extraction are based on the method of context window size and the method based on syntactic analysis. But both the two methods have some shortcomings when being used alone [25-26]. For the method of context window size, if there are less frequent relationship pairs when the context varieties are small and the norms of context vectors are too short, it is more difficult to reliably classify the relation. And for the method of based on syntactic analysis, due to the complexity of text semantics, more linguistic features need to be considered and the complexity of the algorithm is high. Moreover, it is limited to the incomplete corpus dictionary, so the semantic analysis is difficult to use effective dictionary.

In this paper, the relational feature words are extracted by combining the "context window" and the syntactic analysis method, and we propose new rules for syntactic analysis.

Extract the feature word based on the "context window". In general, the number of words in front of the first entity in the sentence, the middle of the two entities, and words behind the second entity are called a "context window". And the specific content contained in the context window is called the context of the two entities. The context of the two entities always have a lot information which can describe the entities relation.

Che proved that in the application of their extraction of the two machine learning methods, when the window size is set to 2, the final result is the best. Therefore, the relational feature word extraction window used in this paper is set to 2.

Extract feature words based on syntax analysis. Some researchers have found that the extraction of features can also be based on syntax and semantics, and this method can improve the performance of the entity relation extraction in some ways.

When applying syntactic analysis to obtain relational feature words, the words on the shortest dependent path of the two entities are usually used as the relational feature words of the entities pair.

However, in reality, not all words on the shortest path between the two entities can provide some information for the entity relation. Moreover, in the syntax analysis, because the Chinese grammar is complex and changeable, analyzing sentences may be wrong and produce noise or omission [27-29].

Therefore, on the basis of the traditional syntactic analysis method and the grammatical characteristics of Chinese, this paper proposes a combined feature extraction algorithm.

The rules are as follows:

- If two entities in the entity pair appear on the side of the core verb, the relational feature of the other side of the core verb in the sentence can be ignored.
- If the entity is centered on the subject, one is the subject, and the other appears the subject of the prepositional phrase, then the verb or noun in the prepositional phrase is the relational feature word.
- When the entities in the sentence appear on the dependent path simultaneously with the two entities related to the subject-predicate relationship (SBV) and the dynamic relations (VOB), then the two dependent paths are extracted as a feature word.
- If the entity has multiple verbs on the dependent path, chose the nearest one.
- Extract nouns or verbs that are directly dependent on entities as the feature words.

We use these rules to extract the feature words, then integrate these words with the first method. It is suitable for our extracting method.

Therefore, the combination of the two methods, we can get a better extraction effect. After the deduplication, we get the set of relational feature words.

2.4 Relational Feature Words Clustering

After extracting the relational feature words of the related entity pairs, we need to effectively cluster all the feature words. Clustering algorithm is a kind of data description method based on the existence of several groups in the whole data set, and the points in each subset have high internal similarity. Then select a representative subset of each cluster to form a subset of features and remove irrelevant and redundant features.

Nowadays, there are lots of the clustering algorithm for studying entity relationship [30-31]. However, for some specific applications, the selection of clustering algorithms depends on the type of data and the purpose of clustering. K-means is one of the very classical clustering algorithms in the partition method. Because of the high efficiency of the algorithm, it is widely used in the clustering of large scale data. Many algorithms are extended and improved around this algorithm. K-means algorithm takes k as the parameter and divide n objects into k clusters to make the internal similarity of the cluster is high while the similarity between the clusters is low. AP cluster algorithm is based on information transfer between data points. It is different from k-means or k center algorithm that AP algorithm does not need to determine the number of clustering before operation. AP algorithm has a definite center of mass rather than the average of k-means is the center of mass. AP algorithm has better robust and higher accuracy than k-means algorithm.

In this paper, the AP clustering algorithm is used to cluster the relational feature words to obtain the words that can describe the entity relation. AP algorithm identifies exemplars among data points and forms clusters of data points around these exemplars. The AP clustering algorithm has the advantages that not need to formulate the final clustering numbers and is not sensitive to the initial value, so the result has a smaller variance error than the k-centers clustering method. It operates by simultaneously considering all data point as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters emerges. It is suitable to entity relationship extraction cluster with advanced clustering capability.

In the process of implementing the AP algorithm, all the data points are regarded as potential examples, which are the nodes of the network. The node uses a recursive way to pass a message, which refers to the similarity of a data point to another data point to find better examples and categories. In this way, we solve the problem that the k-means algorithm is more sensitive to the initial value of the initial clustering center.

The AP algorithm runs as Table 1.

Table 1. Procedure of AP clustering algorithm

AP Clustering Algorithm

Input data: get the cosine similarity matrix $S_{N\times N}$ obtained by using word vector.
Output data: the category after clusteringand the center data node in each category.
Steps:

1. Initialize the matrix A, set $A_{N\times N}$ to 0.
2. Repeat 2.1, 2.2, 2.3, until the sample node for all data points no longer changes, or the number of iteration is maximized.

   2.1 Update any R(x,y) of the R matrix:

$$A(x,y) \leftarrow s(x,y) - \max_{k' s.t. k' \neq k} \{A(x,y') + s(x,y')\}$$

   2.2 Update any A(x,y) of the A matrix:

$$a(x,y) \leftarrow \min\left\{0, r(y,y) + \sum_{x' s.t. x' \notin \{x,y\}} \max\{0, r(x,y)\}\right\}$$

$$A(x,y) \leftarrow \sum_{i' s.t. i' \neq k} \max\{0, R(x',y)\}$$

   2.3 The example node of the data node x is the data node that can maximize the value of R (x, y) + A (x, y).
3. Classify the data nodes with the same example node as the same category, and take the example node as the category word for that category.

By using the above algorithm, we can effectively cluster the feature words and get the descriptive word of each category.

2.5 Evaluation Index of Entity Relationship Extraction

Entity relationship extraction is often evaluated with precision, recall, and F value. Precision and recall rate are two measures widely used in information retrieval and statistical classification to evaluate the quality of results. The precision is the ratio of the number of documents retrieved and the total number of documents retrieved, which measures the accuracy of the retrieval system. Recall rate refers to the ratio of the number of related documents retrieved and the number of relevant documents in the document library, which measures the total recall rate of the retrieval system, and the F value is the harmonic mean of the correct rate and recall rate. And their calculation expressions are as follows:

$$\text{precision} = \frac{\textit{The number of relationship instances that are correctly classified}}{\textit{the total number of relationship instances in a class of test sets}}$$

$$\text{recall} = \frac{\textit{The number of relationship instances that are correctly classified}}{\textit{the total number of relationship instances in a class of test sets}}$$

$$\text{F1} = \frac{2 * \textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}$$

3 Experiments and Result Discussion

In this paper, two parts of the experiment are carried out: the first experiment is based on extracted feature words, and the second the experiment is based on an example of entity relation extraction.

(1) Experiment 1: Analysis on the extraction effect of relational characteristic words. We use the network news corpus (about 1.4G) provided by Sogou Lab as experimental data, and randomly select 50 pairs of entities that appear in the same sentence. According to the extraction method of feature words in this paper, then we compare it with artificial annotated feature words.

The experimental data are as Table 2.

Table 2. Results of Relational Feature Words Extraction

	N-N	N-G	N-O	G-G	G-O	O-O	Total
Number of entities	9	9	9	7	8	8	50
Feature words extracted	9045	5511	19792	11973	5771	938	53030
Artificially marked feature words	1204	443	1817	1850	574	98	5986
Accuracy rate	12.2%	7.8%	8.9%	6.7%	9.3%	7.8%	8.9%
Recall rate	91.8%	97.3%	96.9%	43.4%	93.4%	74.5%	78.7%
F value	21.6%	14.5%	16.3%	11.6%	16.9%	14.1%	16.0%

From the above table we can see, the recall rate of relational feature words is mostly higher than 90%, which means that by using the feature words extraction method proposed in this paper, we can effectively extract the feature words that can describe the entity correctly.

In the extraction method, the accuracy rate and F value are lower, which indicates that the extracted feature words contain more useless words or lower frequency words. But the main purpose of the relational feature words extraction part is to extract the correct feature words as much as possible, and the noise contained in this part will be effectively filtered out at the time of clustering. Thus, although this reduces the overall computational efficiency, this does not impact much on the final output of the entity relationship extraction.

In the result, the feature extraction of the "GEOG-GEOG" entities is worse than that of other entities. The main reason is that there are often no clear relationship words between the geographic entities, which is mainly due to the shortcomings of the "undefined relationship type" of the unsupervised relation extraction.

The above results are the experimental results obtained from the window and syntactic analysis used in this paper. We use it to compare the final results with only the use of the window to extract the feature words.

The experimental data are as Fig. 3.

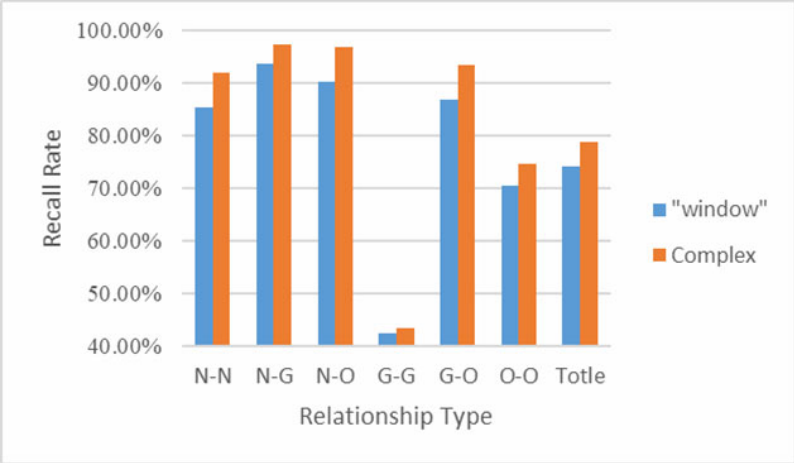

Fig. 3. Comparison of two keyword extraction methods

From the above figure we can see, the comprehensive relational feature extraction method used in this paper has a significant effect on the extraction of feature words. In the results, the lifting effect of NAME-NAME, NAME-ORG and GEOR-ORG was the most obvious, followed by NAME-GEOR, ORG-ORG, then GEOR-GEOR was less effective. This situation is due to that in Chinese grammar when describing the relationship between the names of entities and other entities the word is more direct, and other entities between the relationships between the uses of the word is less obvious.

All in all, by using the method of this paper, it can effectively improve the extraction of feature words, thus it contributes to the extraction of entity relations.

(2) Experiment 2: we extract the relationship with a specific named entity pair. Select the entity pair "Li Xiaolu, Dong Xuan" in experiment 1. By using the feature words extraction method of this paper, we get 112 relational feature words.

The results of clustering are as Fig. 4.

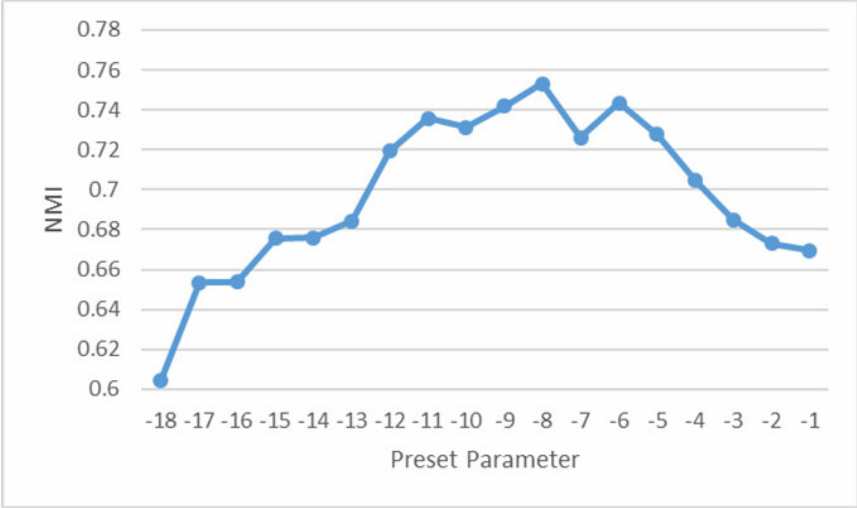

Fig. 4. NMI varies with the preset parameter of AP algorithm

When the preset parameter is set to -8, the NMI value reaches a maximum of 0.7531, which means the effect of clustering at this time is the best. The data entered at this time is divided into 22 categories. Among them, the proportion of the category "Friends" belongs to is much more than other categories, and "Friends" is the example node of the category.

We can get "Friends" is the most able to describe the relationship of the two entities, then we can get the relation extraction result "Li Xiaolu, Dong Xuan, Friends".

4 Conclusion

This paper presents an improved method based on unsupervised entity relation extraction model. Through the process of entity relationship determining, we can select the entity pairs with relationships. Based on the results of the filtering, we construct the extraction rules of relational feature words. Combining the "window" method and syntax analysis method we can get the list of feature words. Then we can use the AP clustering algorithm to cluster the words in the list and get the word that most able to describe the relationship of the two entities.

In the future, we will try to study the work of improving the accuracy of the unsupervised relationship extraction model to improve the computational efficiency of the system.

References

[1] R.C. Bunescu, J.M. Raymond, A shortest path dependency kernel for relation extraction, in: Proc. the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005.

[2] A. Culotta, S. Jeffrey, Dependency tree kernels for relation extraction, in: Proc. the 42nd Annual Meeting on Association for Computational Linguistics, 2004.

[3] J. Jiang, C. Zhai, A systematic exploration of the feature space for relation extraction, in: Proc. Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, 2007.

[4] S. Miller, H. Fox, L. Ramshaw, R. Weischedel, A novel use of statistical parsing to extract information from text, in: Proc. the 1st North American chapter of the Association for Computational Linguistics Conference, 2000.

[5] A. Blum, M. Tom, Combining labeled and unlabeled data with co-training, in: Proc. the Eleventh Annual Conference on Computational Learning Theory, 1998.

[6] T. Hasegawa, S. Satoshi, G. Ralph, Discovering relations among named entities from large corpora, in: Proc. the 42nd Annual Meeting on Association for Computational Linguistics, 2004.

[7] W. Jin, J. Chen, X. Gu, Exploiting web features in Chinese relation extraction, in: Proc. 2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE), 2012.

[8] N. Kambhatla, Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations, in: Proc. the ACL 2004 on Interactive Poster and Demonstration Sessions, 2004.

[9] S. Zhao, G. Ralph, Extracting relations with integrated information using kernel methods, in: Proc. 43rd Annual Meeting on Association for Computational Linguistics, 2005.

[10] G. Zhao, Exploring various knowledge in relation extraction, in: Proc. the 43rd Annual Meeting on Association for Computational Linguistics, 2005.

[11] Y. Liu, K. Xia, J. Zhao, A SVM-based IDS alarms filtering method, International Journal of Security and Its Applications 8(5)(2014) 227-242.

[12] M. Surdeanu, Multi-instance multi-label learning for relation extraction, in: Proc. the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012.

[13] S.-P. Choi, S. Lee, H. Jung, S.K. Song, An intensive case study on kernel-based relation extraction, Multimedia Tools and Applications 71(2)(2014) 741-767.

[14] N. Bach, B. Sameer, A review of relation extraction, literature review for language and statistics II, in: Proc. Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2007.

[15] D. Zelenko, A. Chinatsu, R. Anthony, Kernel methods for relation extraction, Journal of Machine Learning Research, 3(2003) 1083-1106.

[16] S. Sekine, S. Kiyoshi, N. Chikashi, Extended Named Entity Hierarchy, LREC, 2002.

[17] S. Brin, Extracting patterns and relations from the world wide web, in: Proc. International Workshop on The World Wide Web and Databases, 1998.

[18] M. Mintz, Distant supervision for relation extraction without labeled data, in: Proc. the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, 2009.

[19] O. Etzioni, Open information extraction from the web, Communications of the ACM 51(12)(2008) 68-74.

[20] Y. Park, K. Sangwoo, S. Jungyun, Named entity recognition using wikipedia and abbreviation generation, in: Proc. 2014 International Conference on Big Data and Smart Computing (BIGCOMP), 2014.

[21] J. Lehmann, DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia, Semantic Web 6(2)(2015) 167-195.

[22] J. Lafferty, M. Andrew, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proc. the 18th International Conference on Machine Learning 2001 (ICML 2001), 2001.

[23] C. Sutton, M. Andrew, An introduction to conditional random fields, Foundations and Trends® in Machine Learning 4(4)(2012) 267-373.

[24] Y. Ling, Y. Jing, H. Liang, Chinese organization name recognition based on multiple features, in: Proc. Pacific-Asia Workshop on Intelligence and Security Informatics, 2012.

[25] L. Ratinov, R. Dan, Design challenges and misconceptions in named entity recognition, in: Proc. the Thirteenth Conference on Computational Natural Language Learning, 2009.

[26] M. Zhang, J. Su, D. Wang, Discovering relations between named entities from a large raw corpus using tree similarity-based clustering, in: Proc. International Conference on Natural Language Processing, 2005.

[27] S. Zhao, G. Ralph, Extracting relations with integrated information using kernel methods, in: Proc. the 43rd Annual Meeting on Association for Computational Linguistics, 2005.

[28] L. Kebin, L. Fang, L. Lei, Implementation of a kernel-based chinese relation extraction system, Journal of Computer Research and Development 44(8)(2007) 1406-1411.

[29] J. Chen, D. Ji, C.L. Tan, Z. Niu, Unsupervised feature selection for relation extraction, in: Proc. Companion Volume to the Conference including Posters/Demos and Tutorial Abstracts, 2005.

[30] R. Guan, X. Shi, M. Marchese, C. Yang, Y. Liang, Text clustering with seeds affinity propagation, IEEE Transactions on Knowledge and Data Engineering 23(4)(2011) 627-637.

[31] F. Shang, L.C. Jiao, J. Shi, F. Wang, M. Gong, Fast affinity propagation clustering: a multilevel approach, Pattern recognition 45(1)(2012) 474-486.