# The Overview of Entity Relation Extraction Methods

Xian-Yi Cheng, Xiao-hong Chen, and Jin Hua

School of Computer Science and Technology Nantong University, Nantong 226019, China
xycheng@ntu.edu.cn

**Abstract.** The Information extraction can be defined as the task of extracting information of specified events or facts, and then stored in a database for the users' querying. Only with the correct relationship between the various entities, the database can be correctly store in. Entity relation extraction becomes a key technology of information Extraction system. In this paper, we analyze the status of entity relation extraction method; propose several problems for this field to be solved.

**Keywords:** Text mining, entity relation, information extraction.

## 1 Introduction

In the field of information extraction, entity is the basic information elements of the text, and it is the basis of proper understanding of the text[1]. Narrowly defining, the entity is the concrete or abstract entities in the real world, such as person, organization, company, location, etc. Generally, it's expressed by Unique identifier (proper name), such as person's name, organization's name, company's name, location's name and so on. Broad defining, the entity can also contain the time, the expressions of quantifier. The exact meaning of the entity can only be determined by specific application, for example, in specific application, the address, e-mail, telephone number, ship number, conference name, etc. can be use as named entity.

Relation is seen as the link of two entities within a period of time or space[2]. In the Research of information extraction, relation detection plays a key role in the identification and description of events. Thus, the extraction of semantic relation between entities is an important information extraction in the field of basic research. It's used in many research domains, such as, Information retrieval, question answering, ontology construction, information filtering, machine translation, etc.

If we assume, the main function of information extraction is automatically converted form text into data form, the entity extraction determines various elements of the form, and then the entity relation extraction determines the relative position of these elements in the form.

Before discussing the entity relation extraction, firstly we define what are relations and the classification of relations.

From a mathematical perspective, relation is equal to a subset of the Cartesian product; from a computer perspective, relation is a two- dimensional table; from a logical perspective, relation is more than binary predicate. It is noted that the relation

R. Chen (Ed.): ICICIS 2011, Part I, CCIS 134, pp. 749–754, 2011.

what we discussing does not include function, functional relations, unary predicted, numerical relations, event relations, logical relations, etc.

It is more complex in the classification of relations, form the formal of the relations, it has: binary relations and multi-relations; grammatical relations, semantic relations and pragmatic relations; explicit relations and implicit relations. Form the environment of the relations, it has: web entity relations and plain text entity relations. From the pattern of the relations, it has: pre- defined relations and non-pre-defined relations. In recently research, it always pays attention to binary relations, grammatical relations, explicit relations, web relations and pre-defined relations.

The seven pre-defined entity relations we frequently used giving by

ACE (automatic content extraction, ACE) are: part-whole relations (PART-WHOLE), physical relations(PHYS), generic-affiliation relations (GEN-AFF), Metonymy relations (METONYMY), agent- artifact(ART), organizational affiliation relations (ORG-AFF), personal-social relations (PER-SOC), each category also includes a number of sub-types[3]. HowNet also predefines a number of relationships[4].

## 2 Knowledge-Based Entity Relation Extractions

This method of extraction uses linguistic knowledge, before the implementation of extraction, it constructs a pattern set based on words, speech or semantic, and then stored in database. During the relation extraction, the Pretreatment sentence fragment will try to match with the pattern in the pattern set. IF the match is successful; we can conclude that this sentence fragment has a corresponding relationship property of the pattern.

During using the knowledge-based entity relation extraction method, the most difficult step is the construction of relations pattern. Initially, the construction of relations pattern depended on linguists, they analyzed the corpus related to the extraction task in depth, used the existing linguistic achievements, enumerated every possible expression of relationship, constructed the relation pattern by hand. On the one hand, this method make the period of construct the pattern too long, and make the application cost too high; on the other hand, if the extraction system is used for relation extraction in new fields, the Linguists need to extract features of the new field to re-construct relations pattern. This is very difficult to realize in reality. To solve this problem, several scholars have raised different solutions. Douglas E.[5] proposed FASTUS extraction system in MUC-6, express a variety of domain- dependent rules in a extensible, common mode through the introduction of the "macro" concept. Roman Yangarber et al.[6] proposed Proteus extraction system in MUC-7, the pattern constructing method of relation extraction in this system based on sample generalization.

## 3 Feature-Based Entity Relation Extractions

This method is not need to write knowledge rules by special experts, only need many samples used as training data, construct a classifier by a variety of learning method, express as multi- dimensional feature vector by training samples.

During the processing of entity relation extraction using machine learning method based on the feature vector, the most important aspect is the construction method of the sample feature vector. Only select the appropriate features, it can represent the entities correctly, and then improve the learning effect. The appropriate features of the so-called are the classification-related features; these features have a strong degree of differentiation.

The feature vector is a numerical representation of the instance, that is, the instance is converted into feature vector, among them, $x_i$ is ith element of the n-dimensional feature vector. The purpose of the feature-based machine learning method is for a given set of training data (x1,y1)', (x2,y2)', …, (xn,yn)', Which for the binary classification problem yi∈{1,+1}, learning a classification function f, so that for a given new feature vector xi, f can classify it correctly.

The general method that the entities pair (E1,E2) construct the feature vector in a sentence given in fig.1.

E1.TYPE, E2.TYPE, E1.SUBTYPE, E2.SUBTYPE, Order

$W_{i-w}, W_{i-w-1}, ...., W_{i-1}, W_{i+1}, ..., W_{i+w-1}, W_{i+w}$

$t_{i-w}, t_{i-w-1}, ...., t_{i-1}, t_{i+1}, ..., t_{i+w-1}, t_{i+w}$

$W_{j-w}, W_{j-w-1}, ...., W_{j-1}, W_{j+1}, ..., W_{j+w-1}, W_{j+w}$

$t_{j-w}, t_{j-w-1}, ...., t_{j-1}, t_{j+1}, ..., t_{j+w-1}, t_{j+w}$

**Fig. 1.** The feature vector construction

Among them, E.TYPE is the class for the entity belongs, E.SUBTYPE is the subclass for the entity belongs. Order is the position relationship between the two entities, that is, 0 (E1 at the left side of E2), 1 (E1 at the right side of E2), 2 (E1 contains E2), 3(E2 contains E1). I and j respectively are the location of the two entities appearing one after another, Wk and tk respectively are the Chinese words and the speech in location k.

The vector that constructed by all the properties of the entities pair in a sentence is seen as xi, the classification mark is seen as yi, that constitutes a multi-classification sample (xi, yi). The multiple categories can be classified by the binary classifier, it has "one to many" and "two to two" classification methods, we use the "one to many" method.

## 4 Kernel-Based Entity Relation Extractions

Kernel-based methods can make use of many different forms of data organization to express entity relationship. While calculating the distance between the entities, it can use kernel function other than the inner product of eigenvectors. Any kernel function is implicitly calculating the dot product of the object feature vector in high- dimensional feature space, that is , in many cases, it can calculate their dot product not need to

enumerate all the features. In natural language processing, the typical instances are the subsequence kernels and the parse tree kernels.

Compared with the feature vector based method, the advantage of Kernel-based method is that it can express entity relation more flexible, and it can colligate multi-disciplinary knowledge and information through the kernel function mapping. The kernel function has complex excellent properties, thus the final entity relationship distance can be completed by the kernel methods from many different information sources, improving the accuracy. Zelenko proposed a machine learning method based on kernel function for the relation extraction[7]. He firstly defined the kernel function based on shallow parse expression in the text, and designed an efficient dynamic programming algorithm to calculate the value of kernel function. Secondly, used the support vector machine (SVM) and voting perceptions algorithm respectively to achieve information extraction, the experiments showed that the kernel method has very good performance.

## 5 Hybrid Model-Based Entity Relation Extractions

Though the machine learning system become the mainstream recently, Particularly for simple marking problems, knowledge engineering system (rule-based) in general compared to the standard information extraction systems such as: MUC, ACE, and KDD, their performance is the best.

The advantage of the knowledge engineering system is to use the manual mode to extract entities and entity relationships, model can be understood, and can be improved, but improving the effect of pure machine learning system requires additional training data. The impact of adding additional data quickly becomes very small, while the cost of manual annotation of data increases linearly.

TEG is a hybrid entity relation extraction system[2], it is based on knowledge engineering and machine learning systems together. System is based on SCFG. The grammar rules for extraction are artificial regulations, and the probability is trained by a set of annotations. The disambiguation capability of PCFG makes the knowledge engineering system write simple rules, thus eliminating the artificial workload required.

In addition, the training set scale required is lesser than pure machine learning system required (under the same accuracy). Moreover, the rule- establishing and corpus-annotating can balance with each other.

TEG grammar description is composed by the statement and rules, Rules are mainly follow the classic rules of grammar, can be simply written by symbols [ ] and | , the conterminal in the rule must be declared before using. Some extracted entities required, events and instances can be declared as the output concept. In addition, it also needs to declare two types of terminator: glossary and n-gram.

Glossary is a series of terms that is clear or extracted by a single semantic category introduced from external resources. The instances in glossary are: village, city, state, gene, protein, human surname, job name, etc. Some linguistic concepts such as propositions can be considered as the glossary. Actually, for each term, glossary and the conterminal in the rule are equivalent.

N-gram is more complex. When using in rules, it can be extended to any term. However, the probability of generating a given term is not defined in the rule, but it can be obtained from the training set, and based on the previous or the first few terms. Therefore, the one of possibilities of using the n-gram is that TEG rule is context-sensitive.

## 6 Social Network-Based Entity Relation Extractions

That social networks research shows that social networks are an important feature which is the network shown in the community structure. Numerous studies found that many networks are heterogeneous, which is the nature of social networks. It is not a large number of identical nodes connected randomly, but rather a combination of many types of nodes. There are more connections between nodes in the same type, and different types of connections between nodes are relatively small. So the researchers to meet the same type of nodes and edges between these nodes posed by the sub-graphs are networks group or community (Community), shown in Fig.2.

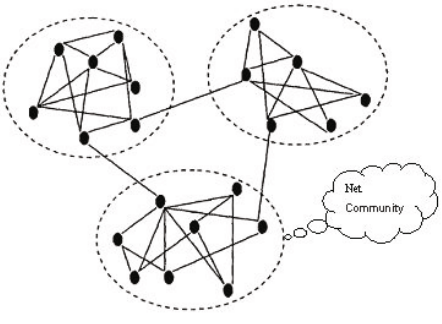


**Fig. 2.** Network of community structure

According to the above characteristics of the community, it can determine the semantic relations in the named entity feature vector and its similarity in structure of the network, the same community named entities with similar characteristics; as expressed by each node represents a semantic feature vector for each of the named semantic entity relationship attribute, so the network of a community has the same semantic relation to a class of named entity pairs. So, just over the network from different communities can be found to achieve the semantic relationship of the named entity clustering[8].

## 7 Other Methods

Skounakis extracted three types of dualistic entity relations from the scientific literature with the model of HHMM[9]. Dan Roth proposed to identify the entity and the entity relation in the sentence with the means of probabilistic framework[10], and fully considered of interdependence between the entity and the entity relationship. In Literature[11], it introduced a method of the whole information methodology which is

used to complete the multi-entity relation extraction, while the clear entity relations and the implied entity relations in the text are extracted at the same time.

## 8 Conclusions

Entity relation extraction has two main ways: knowledge engineering methods and machine learning methods. It needs to summarize manually from the large corpus of knowledge engineering approach and the template system is easy to be in trouble when transplanting, and machine learning methods in the system transplantation have shown very strong advantages, therefore, machine learning methods become a primary way on the research of entity relation. The mixed method reduces the difference from relation extraction system based on the learning and based on knowledge.

## References

1. Zhao, J.-z.: A Research for Semantic Relation Automatic Extraction Among Named Entities in Chinese Professional Domain. Huazhong Normal University (master(s thesis), Shanhai (2007)
2. Feldman, R.: Text Mining Handbook. Cambridge University Press, Cambridge (2007), http://www.cambridge.org/9780512836579
3. http://www.nist.gov/speech/tests/ace/index.htm
4. Liu, Q., Li, S.: Word Similarity Compu-ting Based on Hownet. Computational Linguistics and Chinese Language Processing 7, 59–76 (2002)
5. Appelt, D.E., Hobbs, J.R., Bear, J., et al.: SRI International FASTUS System: MUC-6 Test Results and Analysis. In: Proceedings of the 6th Message Understanding Conference (MUC-6), pp. 237–248 (1995)
6. Yangarber, R., Grishman, R., Tapanainen, P., Huttunen, S.: Unsupervised discovery of scenario-level Patterns for information extraction. In: Proeeedings of the Applied Natural Language Processing Conference (ANLP 2000), Seattle,WA (2000)
7. Grishman, S.Z.R.: Extracting Relations with Integrated Inofrmation Using Kernel Methods. In: ACL (2005)
8. Skounakis, M., Cren, M., Ray, S.: Hierarchieal Hidden Markov Models for Information Extraction. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence, pp. 1010–1018. Morgan Kaufmann, Acapuleo (2003)
9. Dan, R., Wen-tau, Y.: Probabilistic Reasoning for Entity & Relation Recognition. In: 19th international Conference on Computational Linguistics (2002)
10. Zhang, S.: Research on Key Technologies of the Informstion ExtraCtion, vol. 5. Doctoral Dissertation of Beijing University of Posts and Telecommunications, Peiking (2007)