Study of Named Entity Recognition Approaches & Methods

P.N.Santosh Kumar[1], Associate Professor, E mail:pnsk47@gmail.com
Rohith Vedira[2], **K. Sai Akhilesh Reddy**[3]
Dept.of ECM , Srinidhi Institute of Science and Technology, Hyderabad, INDIA

Abstract

This paper explains about Named entity recognition, which is a subtask of information extraction that Recognizes and extracts exact names entities, like Persons names, organizations, locations, time expressions, quantities, monetary values, percentages which are useful for mining information from text. Process of extract names in natural language text is called Named Entity Recognition (NER) task. Correct Named Entity Recognition and Extraction is important to solve Question Answering, Summarization Systems, Information Retrieval, Machine Translation, Video Annotation, Semantic Web Search and Biometrics related problems. Methods like Rule-base NER, Machine Learning-base NER and Hybrid NER are used to identify names from text.

***Key words**: Named Entity Recognition and Extraction, Information Retrieval, Information Extraction, Feature Selection*

1. Introduction

Named entity recognition (NER) involves in different tasks; such a Task is the identification of proper names in Text and another task is the classification of these names into a set of predefined categories of interest, such as person Names (Laxmi, srinu, ramana), organizations, locations (cities names, countries names, rivers names), date and time expressions. The term Named Entity was introduced in the sixth Message understanding conference(MUC-6)

It has provided the benchmark for named entity systems that performed a variety of information extraction tasks. For a machine, recognition of such words in text mining is difficult. The named entities can be classified easily using dictionaries, because most of named entities are proper Nouns, but this is a wrong opinion. Time by Time new proper nouns are created. So, it is not possible to add all those proper nouns to a dictionary. It is difficult to decide their senses even they are added in a dictionary.

2. Ambiguity in NER

Most problems in NER are that they have semantic (sense) ambiguity; a proper noun has different senses according to the context. For illustration, when "Hyderabad **industries**" is an organization, and when it is a location? When "**Tirupathi**" is a person name? and when it is a location name? Or "**Tirupathi visited Lord Balaji temple at Tirumala**", here Tirumala is a location", but in "Tirumala Tirupathi Devasthanams", Tirumala is a part of organization name. Automatically extracting proper names is useful to many problems such as machine translation, information retrieval, question answering and summarization. For instance, the key to a question processor is to identify the asking point (who, what, when, where, etc), so in many cases the asking point corresponds to a Named Entity. In biological text data, the named entity system, can automatically extract

the predefined names like protein and DNA names from raw documents. The goal of named entity recognition and extraction is to extract and classify names into some particular categories from text by respect to the sense of names.

3. Approaches and Methods

Currently automatic named entity recognition and extraction systems have become one of the popular research areas. They can be categorized into four classes; Hand-made NER, Rule-based NER, Machine, Learning-based NER and Hybrid NER.

Hand-made Rule-based focuses on extracting names using human-made rules set. These systems consist of a set of patterns using grammar, syntactic and Orthographic features in combination with dictionaries. An example for this type of system is: "Prime Minister of India Mr. Modhi talks will include discussions on business opportunities in India." In this example a proper noun follows a person's title (Prime Minister), then noun is the name of the Prime Minister and proper noun that is started with capital character (India) after the verb is a Location's name. In this family of approaches, we propose a name identification system based on handcrafted regular expression. They divided the task into three steps: Recognizing Phrases, Recognizing Patterns and Merging incidents However, the rule-based NE systems lack the ability of portability and robustness, and furthermore the high cost of the rule maintains increases even though the data is slightly changed. These type of approaches are often domain and language specific and do not necessarily adapt well to new domains and languages.

In Machine Learning-based NER system, the purpose of Named Entity Recognition approach is converting identification problem into a classification problem and employs a classification statistical model to solve it. In this type of approach, the systems look for patterns and relationships into text to make a model using statistical models and machine learning algorithms. The systems identify and classify nouns into particular classes such as persons, locations, times, etc base on this model, using machine learning algorithms.

There are two types of machine learning model that are use for NER. Supervised and Unsupervised machine learning model. Supervised learning involves using a program that can learn to classify a given set of labeled examples that are made up of the same number of features. Each example is thus represented with respect to the different feature spaces. The learning process is called supervised, because the people who marked up the training examples are teaching the program the right distinctions. The supervised learning approach requires preparing labeled training data to construct a statistical model, but it cannot achieve a good performance without a large amount of training data, because of data sparseness problem. In recent years several statistical methods based on supervised learning method were proposed. A tagging of unknown proper names system with Decision Tree model was proposed. This presented a named entity recognition system based on support vector machines. Unsupervised learning method is another type of machine learning model, where an unsupervised model learns without any feedback. In unsupervised learning, the goal of the program is to build representations from data. These representations can then be used for data compression, classifying, decision making, and other purposes. Unsupervised learning is not a very popular approach for NER and the systems that do use unsupervised learning are usually not completely unsupervised.

In Hybrid NER system, the approach is to combine rule based and machine learning-based methods, and make new methods using strongest points from each method. . Although this type of approach can get better result than

some other approaches, but the weakness of handcraft Rule-base NER remains the same that is when there is a need to change the domain of data.

4. Performance Evaluation

4.1 Definitions and Scopes

Named Entity is a named object of interest such as a person, organization, or location, its task consists of three subtasks namely, entity names, temporal expressions and Number expressions. The expressions to be annotated are unique identifiers of entities (organizations, persons, locations) ENAME, times (dates, times) ETIME, and quantities (monetary values, percentages) ENUME. The task is to identify all instances of the three types of expressions in each text in the test set and to sub categorize these expressions ENAME, ETIME, ENUME.

4.2 Evaluation Metric

The system or method must produce a single, unambiguous output for any relevant string in the text. So evaluation is not based on a view of a pipelined system architecture in which Named Entity Recognition would be completely handled as a preprocess to sentence and discourse analysis. The task requires that the system recognize what a string represents, not just its superficial appearance. Sometimes, the right answer is superficially apparent, as in the case of most, if not all, ENUME expressions, and can be obtained by local pattern matching techniques. In other cases, the right answer is not superficially apparent, as when a single capitalized word could represent the name of a location, person, or organization, and the answer may have to be obtained using techniques that draw information from a larger context or from reference lists.

A scoring model developed for the MUC and Multilingual Entity Task (MET) evaluations measures both precision (*P*) and recall (*R*) Where P= N*umber of correct responses / no of responses,* R = Number *of correct responses/ Number correct in key.* These two measures of performance combine to form one measure of performance, the *F*-measure, which is computed by the uniformly weighted harmonic mean of precision and

$$F= (RP) / 1/2(R+P)$$

The term *response* is used to denote "answer delivered by a name-finder", the term *key* or *key file* is used to denote "an annotated file containing correct answers".

In MUC-7, a correct answer from a name-finder is one where the label and both boundaries are correct. There are three types of labels, each of which use an attribute to specify a particular entity. Label types and the entities they denote are defined as follows:

(i) Entity (ENAME): person, organization, location.

(ii) Time expression (ETIME): date, time.

(iii) Numeric expression (ENUME): money, percent.

A response is half-correct if only the type of the label (and not the attribute) and both boundaries are correct

5. Comparison

We choose some recent efforts with various methods, where all of them use MUC data set. The MUC data collection was derived from the articles of the air-accidents. The performance of the named entity task is measured by three rates, Recall, Precision, and F (β) that were described in the previous section. Some results are shown in the following tables

System	R	P	F (β=1)
NYU	90	93	91.6

Table 1 shows the results of some method that have used Hand-made method. The results show all systems gave high rate in all parameters.

System	R	P	F (β=1)
Association rule Mining	66.34	83.44	70.1
Maximum Entropy	44	61	51

Table 2 indicates results of some systems that have used machine Learning-based methods. The variations in the results were caused by the amount of training datasets and different algorithms.

System	R	P	F (β=1)
NYU	82	91	86

Table 3 report the results of systems using hybrid methods. In these systems gave high rate in all parameters.

6. CONCLUSION

Different types of approaches used for Named Entity Recognition are presented in this paper. All the proposed methods and models have tried to improve precision in recognition module and portability in recognition domain, as mentioned before, one of the most problems and difficulties in NER is to change and switch data domain to new domain and that is called portability. However Hand-make approach can get high rate results in specific domain, still it has problem with broad and new domain. Where Hand-make methods are dependent to domain, Machine learning–based methods is the best independent solution for NER. But high performance of this kind of methods depends on the data training value. This type of approach can get high precision in recognition when amount of data training is huge, and the result is strictly reduce, when data training value is few or malfunction of algorithm.

The Hybrid methods gives good results, but portability of this type of approach is reduced, when they improve precision in recognition by using huge value of fix rules.In the Rule-based method, there was improvement in precision by adding more rules and developing grammatical rules, however portability was reduce automatically, because of fix rules and methods constructors.

7. FUTURE STUDY

Here, we introduce a proposed method in NER, while is a supervised Machine Learning based method by using Support Vector Machine algorithm. The purpose of Named Entity Recognition approach is converting identification problem into a Classification problem and employing a statistical model to solve it. For this approach fuzzy algorithm may applied to improve classification in Support Vector Machines method. In Fuzzy Named Entity Recognition Method the system segments input testing and training data into tokens with a simple tokenizer. Then rich feature sets are selected based on (i) Lexical

information, (ii) Affix, (iii) Previous NE information (iv) Possible NE class & (v) Token feature. We also proposed a new Fuzzy NER and currently implementing the various models to achieve good results and improve our method by defining more token features and a strong fuzzy membership function.

8. References

[1] Message Understanding Conference,

www. nlpir.nist.gov/related_projects/muc.

[2] Y.C. Wu, T.K. Fan, Y.S. Lee, S.J Yen, "Extracting Named Entities Using Support Vector Machines", Spring-Verlag, Berlin Heidelberg, 2006.

[3] I. Budi, S. Bressan, "Association Rules Mining for Name Entity Recognition", Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003.

[4] J. Kim, I. Kang, k. Choi, "Unsupervised Named Entity Classification Models and their Ensembles", Proceedings of the 19th international conference on Computational

linguistics, 2002.

[5] F. Bechet, A. Nasr and F. Genet, "Tagging Unknown Proper Names Using Decision Trees", In proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, 2000.

[6] R. Sirhari, C. Niu, W. Li, "A Hybrid Approach for Named Entity and Sub-Type Tagging" Proceedings of the sixth conference on Applied natural language processing ,Acm Pp. 247 - 254 , 2000.

[7] Collins, Michael and Y. Singer. "Unsupervised models for named entity classification", In proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999.

[8] D.M. Bikel, S. Miller, R. Schwartz, R, Weischedel, "a High- Performance Learning Name-finder", fifth conference on applied natural language processing, PP 194-201, 1998.

[9] A. Borthwick, J. Sterling, E, Agichtein, and R. Grishman, "Exploiting diverse knowledge sources via maximum entropy in named entity recognition", Proceedings of the Sixth workshop on Very Large Corpora, Montreal, Canada, 1998.

[10] A. Borthwick, J. Sterling, E. Agichtein and R. Grishman, "NYU: Description of the MENE Named Entity System as Used in MUC-7", In Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998.