Named Entity Recognition using Statistical Model Approach

Pyari Padmanabhan
Department of Computer
Science and Information
Technology
KMCT College of Engineering,
University of Calicut, Kerala,
India

ABSTRACT

Named Entities (NE) are atomic elements like names of person, places, locations, organizations, quantity etc. Named Entity Recognition is a classification problem. It involves the task of identifying and classifying certain elements in text into predefined categories of named entities. Main subtasks for the Named Entity Recognition involves (1) The Document corpus construction (2) The preprocessing of the documents (3) Determine the contexts (4) Applying the hidden Markov model. In this paper, the hidden Markov model is adopted for the purpose of effective recognition of Named Entities from a document corpus.

General Terms

Named Entity Recognition, Context.

Keywords

Hidden Markov model, preprocessing.

1. INTRODUCTION

Named Entity recognition from a document corpus is a critical task performed especially for researches by Message Understanding Conference (MUC), Multilingual Entity Task Conference (MET), Automatic Content Extraction Program (ACE), and Conference on Computational Natural Language Learning (CoNLL). Named entity recognition (NER) is an important goal in Natural language Processing (NLP). The main aim of NER given a textual document is to determine which atomic elements from the text can be mapped to proper names, such as name of people or places, and recognize the type of the entity like person, location, organization. Identifying the elements or words starting in capital is not always accurate and is insufficient for NER. This is a very complex task due to a variety of reasons.

One is the difficulty in recognizing the nature of the Named Entity, like the instance “Job” in the corpus may not indicate name of a person. Another is the ambiguity that occurs in understanding the same instance. For example, the word “George” may represent a President or may be name of a person. So, a given entity is perceived to be of different types. NER systems have been created that use linguistic grammar-based techniques, statistical models, learning techniques etc.

In this paper, the main goal is to recognize the NE from the document corpus. An initial preprocessing of the text is done. Contexts from the preprocessed data are identified and features like the context weights are used for recognizing the most relevant entity type using the method of hidden Markov model or statistical Markov model that uses novel variables and features related to the instances and its contexts. This paper is organized as following; the section 2 gives an idea of the related works done in the field of NER. The next section describes the steps and methodology involved in recognition. The section 4 gives the experimental results and the final section contributes to the conclusion.

2. RELATED WORKS

Traditional hand-crafted grammar-based systems obtain good precision, but at the cost of lower recall and time. Statistical NER systems require a large amount of manually annotated training data. In methods that use learning techniques as well as combined statistics, measures like number of occurrences of the entity, context frequency, term frequency etc are being considered.

A model based on grammar rules, and statistical models was built for NER by Mikeev et al. [6] without using any lists of named entities or the gazetteers.

Fine-grained proper noun ontology and its use in question answering was told by Mann[7]. The unrestricted text is used and by means of the patterns of occurrence patterns, the question answering function was performed.

Algorithms for named entity classification were used by Collins and Singer [4] including the AdaBoost. The disambiguities, context features and weighted combinations were determined. Etizioni [3] introduced the KNOWITALL system which uses unsupervised technique of extracting large collection of facts from web. B. Favre, F. Béchet, and P. Nocéra [9] introduced the robust named entity extraction from large spoken archives. They considered information extraction from speech input. The output of Automatic speech recognition systems (ASR) was used. They also discussed on the ESTER Named Entity evaluation program.NER with Character-Level Models and a character-Based model was the idea of Dan Klein et. Al.

Weighted logistic regression suggested by Wee Sun Lee, Bing Liu [10], discusses the problem of learning with positive and unlabeled examples by performing logistic regression after weighting the examples. Also, they performed the performance comparisons with SVM model. NER in query was introduced by Hang Li et. al [11] uses the probabilistic approach to the task using query log data & WS-LDA (Weakly Supervised Latent Dirichlet Allocation).D. Nadeau et.al [8] said about generating gazetteers and resolving ambiguity by combining named entity extraction with a simple form of named-entity disambiguation using simple yet highly effective heuristics. A work done for the NER of web documents by Wahiba Ben Abdessalem Karaa [1] used the decision tree approach on the contexts that were recognized.

3. NER APPROACH

The Named Entity Recognition is done in the three main steps that include the Corpus construction, preprocessing and the context and feature recognition. Once the context and its features are identified, the statistical or the hidden Markov model is applied for recognizing the Named entity. The following figure1 shows the simple block diagram for the NER. Each of the steps in the Named Entity Recognition are described below.

3.1 Corpus Construction

For each Named Entity instance also called as the learning example, a set of documents have to be formed. The set of such documents is called as the Corpus. Each document may provide a different clue or feature for the same learning example. Greater the number of documents considered for a same Named Entity instance, more is the chance of getting a better NER performance.

3.2 Preprocessing

Initially the documents are considered which may contain words or elements which are not relevant to be considered for Named Entity recognition. Documents may contain characters like “a”, “as”, “of”, “by” etc. Such irrelevant terms are to be avoided from the document. This is done in the preprocessing. Also, certain other characters like commas, stop characters are avoided. Another important function in this step is to identify the words that start in upper case not only in the starting of a sentence, but also in the middle of a sentence. Such words are more relevant to identify the Named Entity. After preprocessing, the list of relevant words for Named Entity recognition will be available.

3.3 Context and Feature Recognition

A word (or words) that precede or that follow the Named entity is the context. From the document, contexts related to the Named Entities obtained from the preprocessing stage are found. The three words that precede the Named Entity are considered as context.
The features like context frequency and weights are to be found. The number of times the context occurs in the document and corpus is determined, which gives the context frequency. Likewise, for the documents context weight is calculated. The context weight is obtained by determining the term frequency, inverse document frequency measures.

4. NER MODEL

The statistical model or the Hidden Markov model is applied to the context and its features. The possible states that the model passes through are hidden or are unobserved. Each possible state or outcome that is, the named entity is assumed to have a probability.

For any named entity instance, for example, “Sachin Tendulkar”, set of documents or the corpus is trained. Each document related to the instance are preprocessed, the contexts and its features are recognized. These inputs are the parameters for the model. Initially the possible states of a document have probabilities according to the parameters for that document. For a set of documents, the model is applied to get the final output with the maximum probability. The document that recognizes the Named Entity to belong to a named entity class is the most relevant document for that named entity instance.

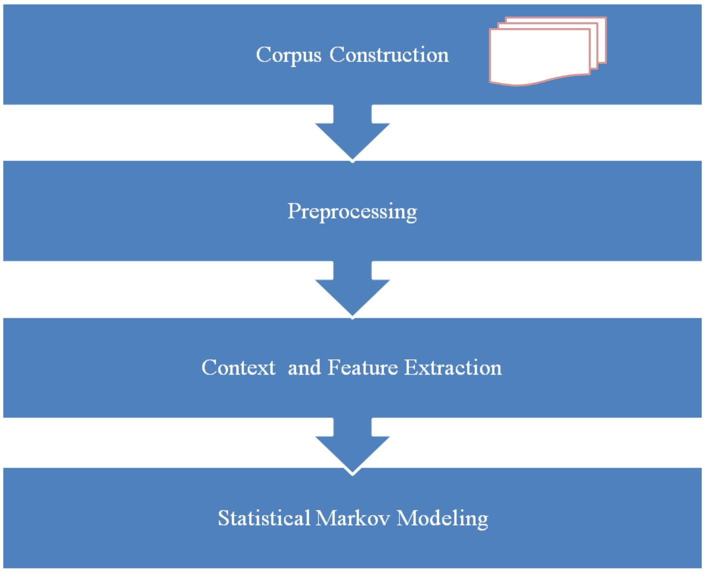

Fig 1: The NER block diagram

5. EXPERIMENTAL RESULTS

Consider a named entity instance “Zidane”. The instance can belong to named entity classes like “person”, “player” etc. So, the named entity class which is most relevant for that instance is to be identified.

A set of documents or corpus for the instance “Zidane” is formed. Each document is analyzed for the instance. The following figure2 represents one such document.

…..**Zinedine Zidane** is a retired **French footballer**. **Zidane played** as an attacking midfielder for the **French national** team, **Juventus and Real Madrid**…………. **Zidane** was named best **European footballer** of the past 50 years by **UEFA**, and has been described by many commentators as one of the greatest players ever…**Zidane won FIFA World Cup** 1998 and **Euro 2000** with **France**………. **Zidane has** won the **FIFA World Player** of the **Year three** times, and **Ballon D'Or** once. **He** was also chosen as the best player of ……… **Zidane retired** from professional football after the 2006 **World Cup**. **He** currently holds the post of **Real Madrid Director** of **Football........**

Fig 2: Example of an entity instance document

The steps in the recognition are as follows:

A number of such documents are trained.

In the preprocessing stage, all the words starting in upper case characters are identified. The words that may occur in upper case in the starting of the sentence but occur in between the sentence in lower case are avoided from the list of named entities. So, after the preprocessing stage a list of named entities that occur in document and are relevant to be considered as named entity are available. The following figure3 is an example of the list obtained.

The contexts or words (2 or 3) that lie to the left or right of the named entities of the list are determined (mostly the words that precede the named entity if it lies in between the sentences).

….
Real
Madrid
Cup
Cup
FIFA
World
Cup
…
….
Real
Real
……

Fig3: Example of the list obtained after preprocessing

For the contexts, the context frequency and weights for the contexts are calculated using the term frequency, inverse document frequency measures. The probability measures are formed for each of the named entity that are in the list. These measures indicate the most relevant named entity for a document of the named entity instance. The following table1 shows example of the relevant named entity for the example document considered. The table indicates “Real” ” Madrid” as the most relevant named entity for the instance.

Table1: Probability values for entities

Entity	**Probability**
Real	0.2800
Cup	0.1500
Madrid	0.2800
FIFA	0.1400
World	0.1500

The probabilities for a set of documents are considered for the same named entity instance “Zidane”. Iteratively, for each document the hidden Markov model conditions are applied. Current state probabilities for the Named entities are considered while the conditions are applied for the next state entities. There are various lists of named entities with different probabilities for each document from which only those named entities that satisfies the parameters of the model and has highest probability at the end is considered to be the most relevant named entity for a particular named entity instance. Finally, the Markov model will recognize a named entity for the instance “Zidane”. The set of the documents considered for the experiment indicated “Footballer” as the most relevant named entity for Zidane and the documents with most probability for “Footballer” as the most relevant document.

The performance of results is measured in terms of recall and precision. The recall and precision values are found to be 73.8 and 81.2 respectively.

6. CONCLUSION

This paper presents the solution to the recognition of the named entities through the hidden Markov model. For this purpose, there is preprocessing stage once the set of documents are trained. The list of named entities for each document is recognized. The context features like the context frequency, context weight, previous state probability values etc form the input parameters to the model. From a number of named entity probabilities, the most relevant one is recognized. The model used has better accuracy. Higher recall and precision values are also observed. The experiments were done by training a modeled corpus. The paper can be extended in future by clustering of the named entities recognized to different classes like “President”, “Footballer”, and so on. Also, the model can be applied in different languages to verify the accuracy.

7. REFERENCES

[1] Wahiba Ben Abdessalem Karaa, Named Entity Recognition Using Web Document Corpus, International Journal Of Managing Information Technology (IJMIT) Vol.3, No.1, February 2011,pp 46-56.

[2] F. Denis, R. Gilleron, and F. Letouzey, Learning from positive and unlabeled examples. *Elsevier. Theoretical Computer Science,* 2005, vol. 348, pp. 70 – 83.

[3] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A.Yates, Unsupervised named-entity extraction from the web: An experimental study, *Artificial Intelligence*, 2005, vol. 65,pp. 91–134.

[4] M. Collins 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, pp.189–196.

[5] C. Krstev, D. Vitas, D. Maurel, M. Tran, Multilingual ontology of proper name, *in Proceedings of the Language and Technology Conference,* pp. 116–119, Poznan, Poland, 2005.

[6] A. Mikheev, M. Moens, and C. Grover, Named Entity Recognition without Gazetteers, *in Proceedings of* Conference of European, Chapter of the Association for Computational Linguistics, EACL '99, pp. 1-8, University of Bergen, Bergen, Norway June 1999.

[7] G.S. Mann, Fine-grained proper noun anthologies for question answering, *International Conference on Computational Linguistics,* COLING-02 on SEMANET: building and using semantic networks, 2002, Vol. 11

[8] D. Nadeau, P. D. Turney, and S. Matwin, Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity, *Lecture Notes in Computer Science, Springer,* 2006, pp. 266–277, Berlin Heidelberg 2006. rs Ltd.

[9] B. Favre, F. Béchet, and P. Nocéra, Robust Named Entity Extraction from Spoken Archives, *in Proceedings of HLT-EMNLP'05,* pp. 491-498, Vancouver, Canada, October 2005.

[10] Wee Sun Lee, Bing LiuLearning with Positive and Unlabeled Examples Using Weighted Logistic Regression *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003),* Washington DC, 2003.

[11] Jiafeng Guo†, Gu Xu‡, Xueqi Cheng†, Hang Li‡, Named Entity Recognition in Query , 2009