

Named Entity Recognition by Using Maximum Entropy

Imran Ahmed and Sathyaraj R

SCSE, VIT University, Vellore-632014
imranahmed2k9@gmail.com, sathyaraj@vit.ac.in

Abstract

Named Entity Recognition (NER) is responsible for extracting and classifying some designators in the given specified text which can be name, location, organization etc. Since the last decade or so, researchers are greatly involved in this area as far as their interests are concerned. It is important procedure to extract the entities in a specified text based on a language which is termed as Natural Language. This language consists of various entities and the collection of such entities is called entity set. These entity sets are maintained in a uniform database called as gazetteer. In this paper we present a methodology called maximum entropy to retrieve the entity sets from the database. The machine is trained in such a way that it will retrieve the words which has the maximum entropy amongst all and has proved to be fastest method to extract and classify the entity sets from the database. The advantages of proposed method include sequence tagging which means this method has increased the freedom of choosing features to represent observations.

Keywords: Name Entity Recognition, Natural Language, entity, entropy, designator, gazetteer

1. Introduction

Name entity is widely used in Information extraction applications. Message understanding Conference is the responsible for information extraction. It also focuses on IE tasks which includes the information extraction used for defense purposes. It was understood that such type of information extraction is very useful and based on some designators this information can be extracted very easily. The designators can be name, location, organization *etc.* The process of extracting such designators from a text is thereby called as Name Entity Recognition. From the last decade, this field has witnessed a lot of research and still going on. NER systems have developed for faster execution of queries [1-12]. The queries means a input which can be a text and output is a information which tells the user which is the name, location or organization present in the text which the user specified as the input.

The aim is to recognize entity sets from a given text .The entity set comprises of name, location and organization. For example ,if the input is given as " John studies in Pune University and lives in California " ,the expected outcome after processing this input text is that it recognizes " John " as a person name , " Pune University " as the organization and " California " as the location. In this paper we have presented maximum entropy method for Name entity recognition.

A lot of work has been done in the field of named entity in recent years which aims to address the ambiguities and the portability issues and some methodology need to be found out to address this problem. The solution demanded large motivation, tractability of the problem and the potential marketability of an accurate named entity system. There were many issues which created a need for such a system which includes more accurate internet search engines and a general document organization. The most general example is of Google, the search in the Google search engine can more optimized if it includes named entity recognizer in it. The second one is the general document organization in

which the user can call up all documents on the organization's intranet which mention a particular individual. Before reading the article a user could see a list of people, locations and companies mentioned in the document.

2. Related Work

The process used in finding out relevant subsets from database is called as feature selection. It contains redundant and irrelevant feature proposed by Lei Yu *et. al.* [2], used in application which contains many features and the user wants to decrease the features. Feature selection has been active field of research since past decade in statistical pattern recognition. The aim of the feature selection is to search the relevant features in the text that matches with the database, if the match is found it retrieves the information otherwise not.

Isabelle Guyon, *et.al.* [3] presented variable and feature selection which has large amount of databases. The areas in which these techniques can be applied include text processing, analysis of genes and chemistry. It provides improved prediction performance of predictors, data retrieval rate has increased to a great extent and the predictors used are cost efficient. Moreover the data which is generated from this process is clear and understood easily.

Asif Ekbal, *et.al.* [4] presented Named Entity Recognition using support vector machine for the classification of individual words into entities and uses various tasks for information retrieval. These tasks include machine translation and system which will report answers when questions are queried. The work reports about the implementation of this project for Hindi and Bengali language using SVM that is support vector machine. It is helpful in predicting the name of person, name of organization and miscellaneous name. There is great amount of improvement which is possible for this project. There is the unevenness in the ratio between names and non-names for Hindi language and moreover this project can be extended for other Indian Languages.

There are many methods for retrieving the information in the literature, one of them is Rule based approach[5]. It uses grammar and grammar based techniques to find named entity tags. The requirement for this work is that it needs rich and expressive rules then only it will give good results. Moreover, knowledge of grammar and other languages is required any rule based system will have following 5 components:

1. A knowledge base which will act as a database itself.
2. An inference engine which will infer information based on the user's input and the knowledge base.
3. **Matching:** As it is using grammar based language then it follows some rule of the language. The statements in this language are called productions. The production has left hand side and right hand side. Now, the left hand sides of all the production are matched against the contents of memory. If the match is found then it is called as "conflict", these conflicts are taken in a separate set.
4. **Conflict-Resolution:** In this stage one of the conflicts in the set which was made in the step 3 is chosen for execution. If none of the productions are satisfied by this conflict then the interpretation is stopped there itself. If the productions are satisfied, then it moves to step 5.
5. **Actions:** In this phase, the actions of the productions selected in the above step are executed. Due to this, the memory contents may be altered. When this stage is finished the execution returns to the first step.

Another model which is used for this purpose is **Hidden Markov Model** proposed by L.E. Baum, *et.al.* [6] which is also known as statistical working model. The system being modeled is Markov process with hidden states. Hidden means the states which are unobserved. In other simple Markov Models like Markov Chain, the observers can see the state and thus the probabilities of state transition are the parameters. Likewise, in hidden

Markov model the state is invisible but the output is visible contrast in case of simple Markov model. The advantage of this model is that it is elegant and easy to understand. Unlike the advantages it has more disadvantages, that is why this model was not applied practically. It is not possible to represent multiple overlapping features and long term dependencies are one of the drawbacks of this model. Another drawback is that this model makes assumptions about the data like Markovian assumptions; it means that the current label depends only on the previous label.

In the class of statistical modeling, there is one model known as Conditional random fields' model [7]. This model is widely used in pattern recognition and machine learning .It is also known as discriminative undirected probabilistic graphical model. **CRF**'s are used mostly used for object recognition and image segmentation. They are undirected graphical models which can be used to calculate the conditional probability of values on assigned output nodes given the values assigned to other assigned input nodes.

There is another classification of models which is the combination of rule based and statistical which is known as hybrid models. **Hybrid NER** [8, 11] approach uses the advantages of both the above mention methods and come out as the strongest method of all.

Various works has been done in the field of Biomedical and Mining Applications for Named Entity Recognition, but it failed in one of the other aspects. Gyorgy Szarvas, *et.al*, [9,13] presented Feature Engineering for Domain Independent NER which says about the searching of information in a text document which is usually in a unstructured format. The author used Text mining [10, 12] for finding the names of relevant entities in the text document. Based on the content of the document evaluation was done and certain actions were taken accordingly. It suffered from less efficiency in retrieving the entities and hence was not that productive.

3. Proposed Work and Implementation

The approach presented in this work is by using maximum entropy method, which uses statistical modeling for turning the notion of futures, histories and features. The maximum possible outcomes or outputs of the model are termed are futures. In this entity sets are selected from the database which is known as gazetteer which has maximum entropy. It is calculated in terms of probability $p(F/H)$, F is the possible futures that can be collected from the space of possible histories H . The amount of random orderings present in the system is called Entropy. The aim objective of this method is to find out such entity set from the database which has maximum entropy with the text specified by the user. So the objective to find such sets that will maximize the probability and that entity is returned to user as an outcome.

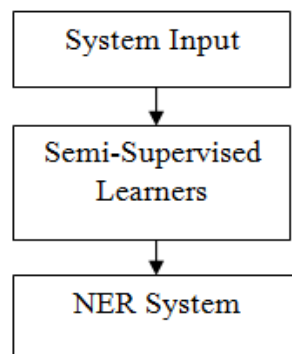


Figure 1. Design of Work

The Figure 1 depicts the architecture of the proposed work. The input to the system is provided by user which is the first stage of the process. The system is trained in a semi-supervised learning methodology. The third step is the information retrieval from the gazetteer.

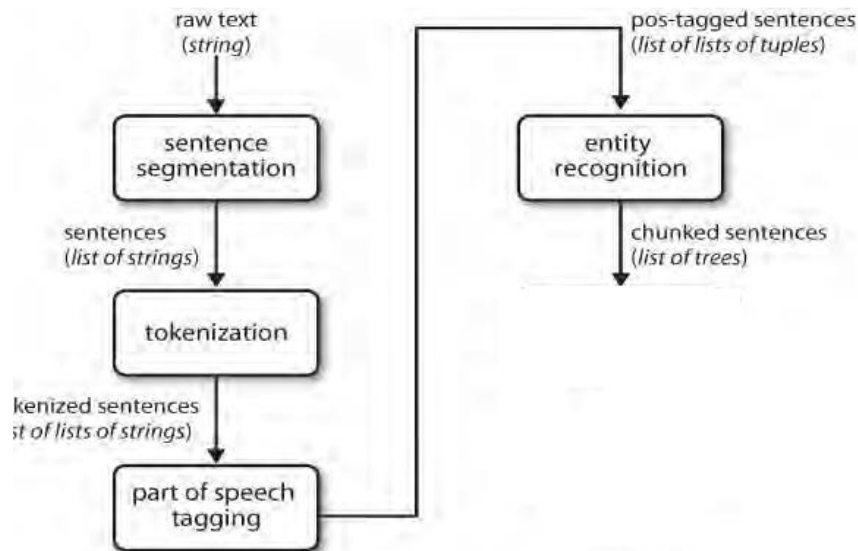


Figure 2. Text Processing of the Input Data

The Figure 2 shows how the text processing of the input data is performed by the system.

3.1. Raw Text

The first module in the operation flow is the raw data which is inputted by the user. It is called as raw data because it is complete sentence in which the name, location entity sets can be in any order. The raw data is in the form of string. Graphical user interface is created for the user to enter the text which the user wants to be checked for named entity recognition.

3.2. Sentence Segmentation

A list is maintained at the back end for storing the separate sentences which is done by the sentence separator, the second module in the work flow. `nltk_sent_tokenize()` function is used to do the sentence segmentation. The inputted text is passed to this function and it returns segmented list of sentences. The parameter which is passed is of type string and the result obtained is also a string.

3.3. Tokenization

The process of splitting the words into tokens on the basis of punctuation and spaces are called Tokenization. The tokens are stored in a list and it resembles each word in a complete sentence which is inputted by the end user. Function used for this purpose is `(nltk.word_tokenize)`, the input sentence is passed as a parameter and it returns the tokens of the sentence. Tokens are actually the individual word in a sentence and it is of type string.

3.4. Parts of Speech Tagging

Next module is tagging of parts of speech and it is done after tokenization. The function of this module is to give parts of speech (POS) tag to each individual token in the user's input. `nltk.pos_tag()` is used for this purpose and token is passed as parameter and the result obtained is a tag.

3.5. Entity Recognition

The last module in the working of this project is the entity recognition. It makes use of maximum entropy method to identify entities in a given text.

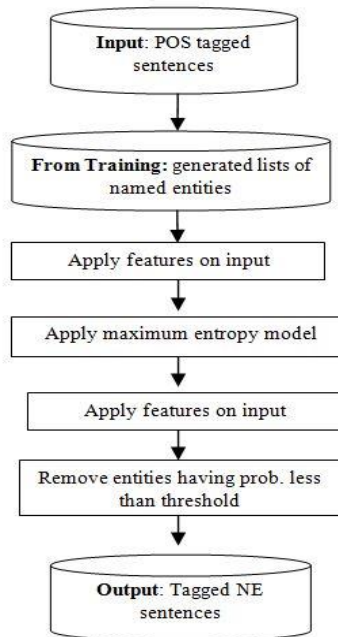


Figure 3. Working Flow Diagram

The Figure 3 depicts the work flow of the algorithm. For optimizing the performance of the system, features which are based on the gazetteer list are used which contains the text files of name which include first name and last name and the location. When the input is taken it is first converted POS tags as described in Section 3.4. Then features are applied to those tags. The maximum entropy model is applied to the intermediate result which is obtained. The entities having less probability than the threshold value are removed as per maximum entropy theory. Finally, we get the output as tagged Named Entity Sentences.

4. Result and Analysis

4.1. Tools Required

The Hardware and the software requirement for the application development are:

1. Windows OS
2. Python 2.7
3. NLTK toolkit
4. Corpus
5. English ACE File
6. Visual studio 2008




Figure 4. Input Form

Figure 4 depicts the input form in which the user has to enter the text for which the user has to apply this algorithm.

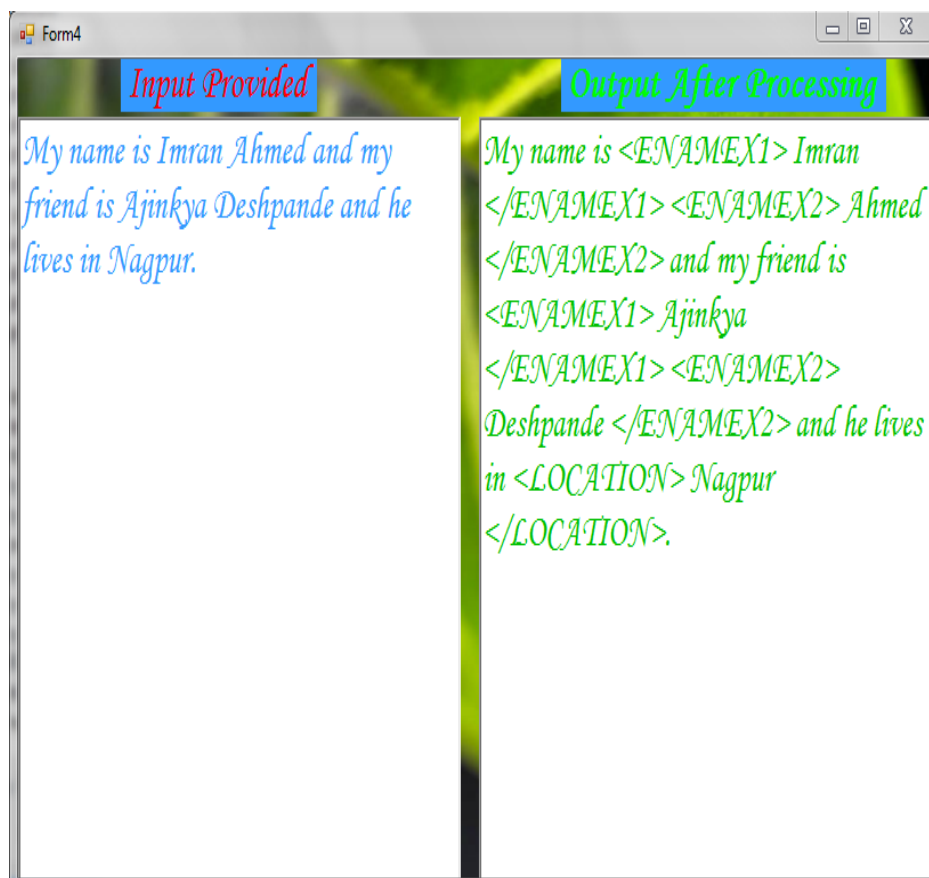


Figure 5. Output Screen

Figure 5 depicts the output of the algorithm of the text which was entered by the user. The name, location are categorized in the form of tags as shown.

4.2 Performance Evaluation

The performance can be evaluated by the terminologies precision, recall and F Measure.

Precision: It is the fraction of the correct answers produced by the algorithm to the total answer produced. The formula for precision is:

$$\text{Precision (P)} = (\text{Corrected answers}/\text{answers produced})$$

Recall: It is fractions of the documents that are matching to the query mentioned and are successfully retrieved. Recall is calculated in the following manner:

$$\text{Recall (R)} = (\text{corrected answers}/\text{total possible answers})$$

F-Measure: It is the harmonic mean of precision and recall. The F-Measure is calculated as:

$$\text{F-Measure} = ((\text{Beta}*\text{Beta}+1)*\text{P}*R)/((\text{Beta}*\text{Beta}*R)+P)$$

	Precision	Recall	F-Measure
OurMethod	98.74 %	93.44 %	89.76
Other Methods	93.21 %	91.55 %	81.35 %

5. Conclusion and Future Work

The work presented is Named Entity Recognition using Maximum Entropy Methodology. It is observed that it provides better results when we are using Maximum Entropy method for retrieving the information from the gazetteer list. The performance of the system is improved than the other methods which were used for retrieving the information as shown in Section 4.

The Future work is to make the automatic training of the system to detect the features with unsupervised approach. Moreover, the work can be extended to build learning base system so that the system can learn from the wrong outputs which will improve the performance of the system.

References

- [1] "Feature Selection Challenges by Neural Information Processing Systems (NIPS) Conference", <http://www.nipsfsc.ecs.soton.ac.uk>, (2003).
- [2] L. Yu and H. Liu, "Efficient Feature Selection through Analysis of Relevance and Redundancy," J. Machine Learning Research, vol. 5, (2004), pp. 1205-1224.
- [3] I. Guyon and A. Elisseeff, "Introduction to Variable & Feature Selection," J. Machine Learning Research, vol. 3, (2003), pp. 1157-1182.
- [4] A. Ekbal, K. Kira and L. A. Rendell, "Practical Approach to Feature Selection," 9th Int'l Conf. Machine Learning, (1992), pp. 249-256.
- [5] R. Farkas, R. S. Gy, "Automatic construction of rule-based ICD-9-CM coding systems. BMC Bioinformatics, vol. 9, no. 3, (2008), [<http://www.biomedcentral.com/1471-2105/9/S3/S10>].
- [6] L. E. Baum, and J. A. Eagon, "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology", Bulletin of the American Mathematical Society, vol. 73, no. 3, (1967), pp. 360—363.

- [7] S. Charles and M. C. Andrew, “An Introduction to Conditional Random Fields for Relational Learning. In: Getoor, Lise; Taskar, Benjamin (Editors.): Introduction to Statistical Relational Learning. MIT Press, (2007) November, Chap. 4, pp. 93-127.
- [8] P. Praveen, “Hybrid Named Entity Recognition System for South-South East Indian Languages”, In Proceedings of IJCNLP workshop on NERSSEAL(Accepted) (2008).
- [9] G. Szarvas, L. Tsai, C. Wu, T. WI, “The maximum entropy approach to biomedical named entity recognition” Data mining in bioinformatics, (2004).
- [10] K. Trentelman, “The Theory Behind the PENG System: Process-able English”, File number: 2009/1016220, Issue date: 2009-06.
- [11] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra, “The Maximum Entropy approach to natural language processes,” Computational Linguistics, vol. 22, (1996), pp. 39–71.
- [12] C. Li and C. Guo, “Survey and taxonomy of feature selection algorithms in intrusion detection system” Inscript 2006, 4318LNCS, (2006), pp. 153–167.
- [13] L.Tsai, C. Wu, T. WI, “The maximum entropy approach to biomedical named entity recognition” Data mining in bioinformatics (2004).
- [14] L. Tan and D. Taniar, “Adaptive estimated maximum-entropy distribution model”, information sciences, vol. 177, (2007), pp. 3110-3128.
- [15] Pandey, “Direct estimation of quintile functions using the maximum entropy principle”, Structural Safety, vol. 22, no. 2, pp.61-79.

Authors



Imran Ahmed, he is currently pursuing post-graduation at VIT University Vellore, Tamil Nadu in Computer Science and Engineering stream. He has done Bachelor of Technology in Computer Science and Engineering from University Institute of Engineering and Technology, CSJM University, Kanpur. His major interest work area is Natural Language Processing, Artificial Intelligence, Cloud Computing and Web development



Sathyaraj R., he is Assistant Professor (Senior) in the School of Computing Science and Engineering at VIT University Vellore, Tamil Nadu. His research interests include software engineering and testing.