A STUDY ON THE APPROACHES OF DEVELOPING A NAMED ENTITY RECOGNITION TOOL

Hridoy Jyoti Mahanta[1]

[1]*Department of Information Technology, Assam University*

Abstract

Named entity recognition (NER) is of vital importance in information extraction in natural language processing. Identifying the named entities in a piece of text and classifying them with proper tagging can help in getting a lot of information engraved in the particular text. The following paper presents brief details about the various approaches in developing a NER. Also an overview of the various models and learning methodologies used for the statistical approach is also provided. The various factors that need to be considered in developing this tool are also stated.

Keywords: *Named Entity Recognition, Information Extraction, Corpus, Enamex.*

----------***----------

1. INTRODUCTION

Natural Language Processing (NLP) is theoretically motivated range of computational technique for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human like language processing for range of tasks or application. [see. Encyclopaedia of Library and Information Science, Liddy, E. D]

Natural language processing has been exploring various fields in language processing through information extraction, information retrieval, machine translation, speech recognition, question answering etc. With growing interests of the scholars, linguist, scientists many new methodologies for language processing have come up.

Named Entity Recognition is a subtask of information extraction that seeks to locate and classify the proper names in a text. NER is an indispensible part in any Natural Language Processing applications such as: question answering, machine translation, information extraction and information retrieval

2. ABOUT NER

With a large focus on information extraction it was found necessary to identify information units such as names of person, place, organization, temporal expressions etc. It became evident that, in order to achieve good performance in information extraction, a system needs to be able to recognize names. So identifying these entities became a prime task of IE and thus "named entity recognition" was acknowledged. Much research has since been carried out on NER, using both knowledge engineering and machine learning approaches.

In its canonical form, the input of an NER system is a text and the output is information on boundaries and types of NEs found in the text. For instance consider the following text "Richard Marx and Julia Smith work for Microsoft Corp. and are currently settled in Boston." The above text fed to an NER will produce an output of the form as shown below "Richard Marx [PERSON] and Julia Smith [PERSON] work for Microsoft Corp.[ORGANIZATION] and are currently settled in Boston [LOCATION]."

3. RELATED FACTORS

As large number of papers related to NER was presented in the message understanding conference (MUC), some basic factors came up, which needed prime emphasis while working in this field.

3.1 Language Factor

A large amount of research work has been done in English. Most of the domains in English have been explored. But this has confined the work to the particular language only. Language independence and multilingual are the prime problems in this field. Languages like German, Spanish and Dutch have been studied in their own prospect. Chinese, Japanese, French, Italian, Greek, have been studied in abundant of literature. Survey on Bulgarian, Hindi, Danish, Korean, and Turkish are in progress. Even Arabic has started receiving attention. But this works continues to be in it own boundary. Developing a multilingual NER is of prime attention in the current scenario. [1]

3.2 Domain Factor

Initial work for NER ignored the two main factors of a language, firstly the type of text or textual genre like informal,

scientific, technical etc and secondly domain like business, sports, tourism etc. The domain and the type of text collectively form the corpus to learn and test the system. But a single domain may have many corpuses within it. Hence a particular language will have large corpora of text to evaluate. Relating the corpora of one domain with another is a major challenge. [1]

3.3 Entity and Tagging

The "Named entity" expression in Named entity recognition confines it to identify only those entities which are rigid and have some particular designation. For English these rigid designated entities are the proper nouns.

The most studied entity types are names of "person", "location" and "organization". These types were collectively called the "Enamex" in MUC-6. However, many papers describe that these types can be further divided to subtypes like person can be a doctor, celebrity, politician and locations can be country, state, heritage sites etc (M. Fleischman 2001, S. Lee & Geunbae Lee 2005). Also there are other entity types beyond the basic three such as time, money, occasions, accessories etc. These entities are collectively classified as miscellaneous type. All the named entities beyond enamex are put in this type. But the miscellaneous type includes a large number of entities collectively which is not efficient. An NER with large number of identification and tagging will be considered as an efficient one. [1]

4. APPROACHES

Development of an NER basically has two approaches Rule Based Approach and Statistical Approach. Another approach, the Hybrid approach can be formed by combining the above two approaches. All these approach's efficiency varies and the one having the maximum efficiency should be followed.

4.1 Rule Based Approach

A rule based system is based on rules given by the linguists and dictionary mapping. The system gives the output by matching the rules and dictionary mapping.

In this approach for each individual classification of named entities different rules are given by the linguist. These rules are then implemented when the user enters the text. Whenever the system gets the text it first searches for the named entity and then compares it with the rules that have been used. Once the rule is matched, the system fetches the classification and gives the required output. [2]

4.2 Statistical Approach

The statistical approach is quite different from the rule based approach. Where in the rule based approach the task of detecting and classifying the named entity solely depended on the rules given by the linguist, in the statistical approach mathematical logic and formulas are used for the same purpose.

Statistical approach differs from the traditional processing in that instead of having a linguist manually construct some model of linguistic phenomenon, the model is instead (semi-) automatically constructed from linguistically annotated resource.

Methods for assigning parts of speech tags to words, categories to texts, parse trees to sentences, and so on, are (semi-) automatically acquired using machine learning techniques.

In statistical approach a corpus is initially studied and based on the corpus a training module is made where the system is trained to identify the named entities and then on their occurrences in the corpus with particular context and class a probability value is counted. Every time when text is given based on the probability value the result is fetched. [3, 4]

4.2.1 Models in Statistical Approach

There are various models that are used for the developing statistical NER. These models have their own mathematical approaches and techniques for training the corpus, determining the probabilistic values and have their own methodologies of working to get the desired result.

4.2.1.1 Hidden Markov Model

A Hidden Markov Model (HMM) is a graphical model which allows us to express conditional probability distributions based on a limited history (the Markov property). There are two types of conditional contexts in a HMM:

a) Observation contexts: Those contexts which are directly related to observations we make.
b) Hidden contexts: Those which are related to hidden (or latent) states.

In the Markov chain, we only see the first type: observation contexts; those contexts which define the conditioning context over each observation in the sequence. The HMM introduces the concept of a hidden state or some state which is not part of the input sequence.

We use the HMM to model a sequential process where we believe the observed sequence is actually dependent on a separate process. In n-gram language modeling, we assume that the observation of a word was dependent on the observation of previous words. But we know that it is not only the words that matter, but also their particular disambiguated usage. In some cases, it is important to know the syntactic category of a word (the part-of speech) in order to predict the next word. This may more important than knowing the actual word.

In a regular Markov Model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a Hidden Markov Model (i.e., HMM), the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; even if the model parameters are known exactly, the model is still 'hidden'. [3]

Applications of HMM

1. Speech Recognition
2. Machine Translation
3. Parts of speech Tagger

4.2.1.2 Maximum Entropy Model

The principle behind maximum entropy model is that subject to known constraints the probability which best represents the current state of knowledge is the one with largest entropy.

A Maximum Entropy model is well-suited for such experiments since it combines diverse forms of contextual information in a principled manner, and does not impose any distributional assumptions on the training data.

The idea of maximum entropy modeling is to choose the probability distribution that has the highest entropy out of those distributions that satisfy a certain set of constraints. The constraints restrict the model to behave in accordance with a set of statistics collected from the training data. In particular, the constraints demand that the expectations of the features for the model match the empirical expectations of the features over the training data. The maximum entropy model has always produced best result as per statistics.

For example, the parts of speech tagging with maximum entropy model by Adwait Ratnaparkhi, University of Pennsylvania (1998) has given an accuracy of around 96.6% in training. [3, 5, 6, 7]

Applications of maximum entropy model:

1. Speech Recognition
2. Machine Translation
3. Parts of speech Tagger
4. Named Entity Recognition

4.2.1.3 Conditional Random Field

A conditional random field (CRF) is a type of discriminative undirected probabilistic model. It is most often used for labeling and parsing of sequential data, such as natural language texts or biological sequence and computer vision. [http:// en.wikipedia.org/wiki/]

A CRF is an undirected graphical model in which each vertex represents a random variable whose distribution is to be inferred. Edges represent dependencies between two random variables. In most models, only pair wise dependencies between variables are modeled.

A CRF is a Markov random field that was trained discriminatively. Therefore it is not necessary to model the distribution over always observed variables, which makes it possible to include arbitrarily complicated features of the observed variables into the model.

CRF may further be classified to two types:

Higher order CRF: The CRF model can be extended into higher order by increasing the number of the sequence of label. In this case the labels are made dependent on fixed numbers of variables. If the number of variable becomes very large then training and inference becomes complex and hence alternative training and inference methods are applied.

Semi Markov CRF: The semi-Markov conditional random field (semi-CRF), models variable-length segmentations of a label sequence. This provides much of the power of higher-order CRFs to model long-range dependencies of the sequence at a reasonable computational cost.

Application of CRF model:

1) Shallow parsing
2) Named Entity Recognition
3) Gene Finding

4.2.2 Learning Methodologies

The statistical approach identifies the named entities through their distinctive features and some mathematical implementations. Unlike early approaches which used handcrafted rules, the statistical approach makes the system learn to identify the named entities through training. There are three main methods to learn the system.

4.2.2.1 Supervised Learning

Supervised learning uses pre-annotated corpus to train the system. This learning method is used by all the models like Hidden Markov, Maximum Entropy and Conditional Random Field model. The system here reads the annotated corpus, memorizes it and used the same to identify the entities from the input text. But for better performance of the system the training corpus should be very large. But unavailability of such large corpus leads to use semi supervised and unsupervised methods. [1].

4.2.2.2 Semi Supervised Learning

In semi supervised learning some initial entities called seeds are trained into the system. The system then searches for these seeds and identifies them. Then the system tries to identify

other entities that appear in similar context where it identified the seeds. The learning process is then again applied using these new contexts. Lexical form, patterns, syntactic analysis forms basic features in semi supervised learning [1, 8].

4.2.2.3 Unsupervised Learning

The main technique behind unsupervised learning is clustering. Here a large number of entities occurring in similar context are grouped into one unit and the system is made to learn this cluster. Whenever it is implemented it looks for the clusters and entities resembling similar contexts are identified as ones in the trained cluster. [1]

4.3 Hybrid Approach

The two traditional approaches of Named Entity Recognition (NER) are: rule based approach and statistical or machine learning approach. The rule based approach achieves better accuracy but requires a large amount of labor by linguist and domain experts.

Because of this, recent research in NER is concentrated in machine learning technique which requires only manually training set documents. Apart from these traditional approaches, the latest approach is the hybrid approach which combines both machine learning techniques and manually written rules. Hence, it benefits from both the approaches and can outperform manually written rules and machine learning. [9]

5. CONCLUSIONS

Named Entity Recognition (NER) is a volatile field in current time. It strives to extract and classify the rigid designators from a piece of text. Many new researches are going on in this field. In this paper, an overview of the two basic approaches to develop a NER has been described. The statistical approach includes details of some of the existing models which are used. Also the various learning methodologies to train the system for this approach are also discussed.

Finally the new approach, the Hybrid approach that can be obtained by combining the two basic approaches has also been briefly stated.

Named Entity Recognition will have a vital impact on our society as it will enable us extract more and more information from a piece of text by identifying and classifying the named entities which may be known or unknown to us.

It is indeed the basis of a major advance in biology and genetics, enabling researchers to search the abundant literature for interactions between named genes and cells.

ACKNOWLEDGEMENTS

Thanks to Dr. Debasri Chakraborty Dubey, Mr. Sahzad Alam, Subhrajyoti Puzari and Pranjal Goswami for their valuable help.

REFERENCES

[1] David Nadeau, Satoshi Sekine, "A survey of named entity recognition and classification", NRCC, New York University, 2007

[2] Kashif Riaz, " Rule-based Named Entity Recognition in Urdu", Proceedings of the 2010 Named Entities Workshop, ACL 2010, July 2010

[3] Christopher D. Manning and Hinrich Schütze, "Foundations of Statistical Natural Language Processing", MIT, May, 1999

[4] Chris Callision-Burch and miles Osborne, "Statistical Natural Language Processing", (A handbook for language Engineers), 24 Feb, 2003

[5] Andrew Borthwick. "A Maximum Entropy approach to Named Entity Recognition." New York University, September, 1999

[6] Hai Leong Chieu and Hwee Tou Ng "Named Entity Recognition with a Maximum Entropy Approach", Proceedings of CoNLL-2003

[7] Adam L. Berger et. al: "A maximum Entity Approach to Natural Language Processing" Computational Linguistics - COLI , vol. 22, no. 1, pp. 39-71, 1996

[8] David Nadeau, "Semi Supervised NER: Learning to recognize 100 entity type with little Supervision", University of Ottawa, 2007

[9] Moshe Fresko, Binyamin Rosenfield et. al, "Hybrid Approach to NER by MEMM and Manual Rules" Proceeding of: International Symposium on Artificial Intelligence and Mathematics (ISAIM 2006), Fort Lauderdale, Florida, USA, January 4-6, 2006