Abstract: Named Entity indicates the names of person, organization, location or anything. The main goal of this paper is to recognize the Named Entity to which any Named Entity instance belongs to. For this purpose, a set of documents related to the named entity instance are collected. Firstly, the named entity instance and few words that precede the instance (called as context) are recognized. Context frequencies, probabilities and weights are calculated. Next, a preprocessing is done on the documents. The words starting in upper case are more probable to belong to a named entity class. Statistical Markov model is applied to each of the documents iteratively to eventually obtain the named entity class for the instance.

Keywords: Context, context weight, Named entity recognition, statistical Markov Model.
1. Introduction:

Named Entity indicates the elements of a document which may indicate the name of any person, organization, location, institution etc. The main aim is to map the elements of a document to proper names and define an entity type for the named entity. Named entity recognition is one of the important tasks in Natural Language Processing (NLP) and is being performed in important conferences like the Message Understanding Conference (MUC). Named entity recognition is an important subtask in information extraction.

In the named entity recognition often more than one recognition is possible for a named entity instance, for example, the word ‘Manmohan Singh’ may belong to a named entity type Prime Minister or Indian. So, for the named entity instance the entity type that is most relevant has to be recognized. Here, the type Prime Minister is more relevant for ‘Manmohan Singh’. Hence, the recognition must be the most relevant for all the set of documents considered. In this paper, first the set of documents called as the corpus is constructed and iteratively each document related to the instance is processed. The words starting in upper case and the contexts are identified. Measures for the contexts like the context frequency, weight and context probabilities are investigated. The hidden Markov model is then applied to the list of named entities observed to eventually obtain the most relevant named entity for the instance.

There are various works done in the field of named entity recognition. The next section describes the important works done in this field. Section 3 gives an idea about the architecture of the proposed system. Section 4 shows the results obtained after implementing the proposed algorithm. Section 5 gives the conclusion about the findings.

2. Existing Works:

There are various methodologies done so far in the named entity recognition area. The initial works are based on the hand crafted rules and grammar. This method is less efficient in terms of the time taken for its implementation. The works were done in identifying the proper name in the documents by applying the procedure of question answering [2]. The use of grammar without maintaining the gazetteers or previous named entity lists was also done [3]. Another work was the use of algorithm for named entity classification which considers the
combination of context frequency and weight to find disambiguities [4]. Later algorithms based on unsupervised learning technique for extracting facts from web were introduced [5].

Works were also done on learning the positive and unlabelled examples by weighting the examples which also carried out performance comparisons to the SVM model [6]. Another work considers the spoken archives from which the named entities are recognized [7]. The use of heuristics and the gazetteers for named entity extraction and disambiguation was also done [8]. Probabilistic methods were used in order to find the named entities from the queries using the query log data [9].

Another work in the named entity recognition considers the construction and training of web documents for a named entity instance. The calculation of the context weights, term frequency and inverse document frequency are done for the tree construction. The recognition involves the use of decision tree for the calculated measures [1].

In this paper also measures for the contexts like the context frequency, context probabilities, context weight are calculated. Based on the values obtained, probabilities are assigned to different named entity classes which are used in the hidden Markov model. The main performance measures are recall and precision. Recall indicates the number of correct named entity found by the system over the total number of correct named entity in the document corpus and precision indicates the number of correct named entity found by the system over the total number of named entity found in the corpus.

3. Proposed Architecture:

For recognizing the named entity, consider a named entity instance. The recognition process undergoes the following stages. The figure1 shows the diagrammatic representation of the architecture.

3.2. Collecting the set of documents

For the named entity instance, a number of related documents are collected. The set of documents also called corpus is further analyzed. From these documents, a number of named entities are possible for a single instance. Hence, the most relevant named entity has to be found. The main issue faced here is the ambiguity among the named entities that is; any
instance should belong to only one particular class. The following steps are followed in order to disambiguate among the named entities.

3.3. Processing of the Documents:

The documents are processed before the named entity recognition. There may be words in the documents which are not related to the named entities. Words of shorter length and of lesser number of characters, certain words that are common in most of the documents, some special characters like commas, stop character, colon etc are eliminated from the document. The resulting document is more precise and there is more probability for efficient recognition of named entity.

3.4. Formation of input for the Markovian model:

From the documents, the words are recognized which are either the named entity instance or which start in upper case. The words of the documents are to be identified according to their priority or relevance. The words that mark the beginning of a sentence start in upper case. Most importantly, there may be words that lie in between sentences in the corpus which begin in the upper case. There is a chance that these words are more related to the instance and have a higher priority. Also there may be words in upper case whose priority increases as its distance increases from the named entity instance for a sentence.

After identifying the words according to their priority in the document, context that precede these words are determined. Further, the number of times each context repeats in the corpus that is the context frequency is found. Also, the probabilities of each context and the related context weights are evaluated. Accordingly, a list of probable named entities is found for every document. This forms the input to the next stage where the hidden Markov model is being applied.

3.5. The Markovian model

In the implementation of the Markovian model, initially a document is analyzed through all of the above steps and the list of the named entities form the input. Each of the probable named entity is considered in the model. Iteratively evaluate the probabilities for all the documents in the corpus. For each document, the named entity probabilities are evaluated and compared based on certain conditions. The probabilities keep varying for named entities after analyzing a number of documents one by one. But certain named entities have their
probabilities consistent and hence after a definite number of documents considered, highly probable consistent named entity is recognized. So, the iterative approach on the documents based on certain conditions on the context frequencies, weights and their probabilities result in the most relevant named entity.

**Figure 1: The NER Architecture**

4. Recognition Model and Results:

For experimentation, consider a named entity instance “Manmohan Singh”. This instance has to be recognized into relevant named entity instance. For recognition, the following steps are done:

- First, collect the corpus related to “Manmohan Singh”. A set of documents forms the corpus. The following figure 2 shows an example document.

```
………………..India’s fourteenth **Prime Minister, Dr. Manmohan Singh** is rightly acclaimed as a thinker and a scholar. He is well regarded for his diligence and his academic approach to work, as well as his accessibility and his unassuming …………………… **Prime Minister Mannohsan Singh** was born on September 26, 1932, in a village in the Punjab province of undivided India……….. Dr. Singh completed his Matriculation examinations from the Punjab University in 1948…………
```

**Figure 2: Example of a document**
✔ Process all the documents to eliminate the words and characters of small word length and which are unimportant. A precise relevant corpus is then formed.

✔ Form the context which includes the 3 or 4 words that precede the instance. Also find the upper case starting words which may indicate the named entity. A list of possible words is obtained. Find the number of times the terms repeat that is the frequency, also the weight of the words in the list which is considered to increase with increase in the distance of word from the instance. Accordingly a list of most probable named entities is obtained with their probable values. The following figure 3 shows an example of the list.

![Example of named entity list](image)

*Figure 3: The example of named entity list*

✔ A list along with their probability values, the frequency measures, weight values of a single document are applied to the Markov model. Iteratively for all documents, compare the values of probabilities based on certain conditions to finally obtain the most apt named entity for “Manmohan Singh”. The following figure 4 shows an example after few iterations of the hidden Markov model.

<table>
<thead>
<tr>
<th>Entity</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prime</td>
<td>0.2800</td>
</tr>
</tbody>
</table>
Prime Minister is the named entity observed with the highest consistent probability value for the instance after a number of documents are analyzed. Hence “Manmohan Singh” instance belongs to the class “Prime Minister”.

The results show favorable performance measures. The recall value is evaluated to be 73.8 and the precision value to be 81.2.

5. Conclusion:

This paper gives an efficient method of recognizing named entities. An iterative process is done on the corpus. The corpus here is a trained corpus for an instance. The named entity recognition is a classification problem. Here the named entity instances are classified into their named entity classes. The result gives satisfying performance measures like the recall and precision. As an extension, the same technique can be performed for various instances of different languages, not only English. Also, other features for the document can be identified to improve the system. The clustering of named entities is also an area of future work.
References:


