Exploiting Wikipedia in Identifying Named Entities: A Language-Independent Approach

by

Mahathi Bhagathula, GSK Santosh, Vasudeva Varma

in

The 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)

sheraton, Maui, USA

Report No: IIIT/TR/2012/-1

Centre for Search and Information Extraction Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
October 2012
Exploiting Wikipedia in Identifying Named Entities: A Language-Independent Approach

Mahathi Bhagavatula
Search and Information
Extraction and Retrieval Lab, IIIT Hyderabad
mahathi.b@research.iiit.ac.in

Santosh GSK
Search and Information
Extraction and Retrieval Lab, IIIT Hyderabad
santosh.gsk@research.iiit.ac.in

Vasudeva Varma
Search and Information
Extraction and Retrieval Lab, IIIT Hyderabad
vv@iiit.ac.in

ABSTRACT

This paper details the approach to identify Named Entities (NEs) from a large non-English corpus and associate them with appropriate tags, requiring minimal human intervention and no linguistic expertise. The main objective in this paper is to focus on Indian languages like Telugu, Hindi, Tamil, Marathi, etc., which are considered to be resource-poor languages when compared to English. The inherent structure of Wikipedia was exploited in developing an efficient co-occurrence frequency based NE identification algorithm for Indian Languages. We describe the methods by which English Wikipedia data can be used to bootstrap the identification of NEs in other languages. On a dataset of 2,622 Marathi Wikipedia articles, with around 10,000 NEs manually tagged, an F-Measure of 81.25% was achieved by our system without availing language expertise. Similarly, an F-measure of 80.42% was achieved on around 12,000 NEs tagged within 2,935 Hindi Wikipedia articles.

1. INTRODUCTION AND RELATED WORK

Named entity recognition (NER) (also known as entity identification and entity extraction) is one of the important subtasks of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, etc. NER has many applications in NLP, e.g., data classification, question answering, cross language information access, machine translation system, etc. Unlike in the case of Indian Languages, a lot of work has been done in the field of NER for English language. The existing approaches can be classified broadly into Language-Dependent and Language-Independent approaches.

Language-Dependent Approaches:

A number of grammar rules can be crafted to build a NER system for a given language (?! ? ??). However, the grammar-based techniques require linguistic expertise and strenuous efforts to build a NER system for every new language. Moreover, for English and many European languages, capitalization is a major clue for crafting grammar rules. Whereas, Indian languages do not have such feature. Hence, crafting rules becomes an arduous task for Indian languages and can be safely avoided when there is a requirement to build a generic NE system for several Indian languages.

Language-Independent Approaches:

Language-Independent techniques developed for the task of NER are predominantly statistical approaches. Statistical NER systems typically require a large amount of manually annotated training data. Several Machine Learning techniques had been successfully used for the NER task of which Conditional Random Fields (CRF) (?), Hidden Markov Model (HMM) (? ??), Maximum Entropy Model (MEMM) (? ??) were prevalent. However, with the serious lack of such manually annotated training data, the task of high-performance statistical NER system projects as a major challenge for Indian languages.

As an alternative, research has moved towards techniques which involve Wikipedia as a major source of data. Wikipedia has been the subject of a considerable amount of research in recent years. The most relevant work to this paper are ? ??.

? used Wikipedia, particularly the first sentence of each article, to create lists of entities. Rather than building entity dictionaries, associating words and phrases to the classical NE tags (Person, Location, etc.), they used a noun phrase following the verb forms 'to be' to derive a label. For example, they used the sentence 'Franz Fischler ... is an Austrian politician' to associate the label 'politician' to the surface form 'Franz Fischler'. They proceeded to show that the dictionaries generated by their method are useful when integrated into an NER system. It is to be noted that their technique relied upon a part-of-speech tagger.

? emphasized on the use of links between articles of different languages, specifically between English Wikipedia (the largest and densely linked) and other languages. Their approach used English Wikipedia structure namely categories and hyperlinks to get NEs and then used language specific tools to derive multilingual NEs.

However, the existing techniques suffer from extensibility to resource poor languages. Few major limitations we have observed are:

1. At least one of the language resources or tools like
dictionaries, POS taggers, gazetteer lists, etc., were used for constructing a NER system. Such resources are limited across Indian languages.

2. In the absence of sufficient manually annotated data, the statistical techniques don’t promise commendable results.

3. They used either the first sentence, title, or category information from Wikipedia articles. However, there lies a scope for wider exploitation of many other structural characteristics of Wikipedia.

In this paper, we aim to overcome the above limitations. This paper focuses on building a generic-purpose NE identification system for Indian languages. Given the constraints for resource-poor languages, we restrain from developing a regular NE recognition system.

The contributions of this paper include: developing a language independent system without using any language-dependent tools or resources such that it can be extended to any other language, exploiting several structural aspects of Wikipedia and proposing a co-occurrence frequency based approach to map similar content across languages which has the potential to be applied elsewhere (explained in detail in 3.4.2).

2. WIKIPEDIA STRUCTURE:

   For remainder of the paper, we refer to the following format of the dataset. Within Wikipedia, we have availed four major structures:

   **Category links**: Links from an article to a special ‘Category’ pages, represented in the form of [[Category:One Day Internationals]], [[Category:International Matches]], {{Sports}}. The first two are direct links to Category pages. The third is a link to a Template, which links the article to ‘Category: Sports’. We will typically say that a given article belongs to all these category pages.

   **InterLanguage links**: Links from an article to a presumably equivalent article in another language. For example, in the Hindi language article ‘Mahendra Singh Dhoni’ one finds a set of links including [[en:Mahendra Singh Dhoni]] and [[mr:महेंद्र सिंह धोनी]]. These represent links to English and Marathi language articles on ‘Mahendra Singh Dhoni’, respectively. In almost all cases, the articles linked in this manner represent articles on the same subject.

   **Subtitles of the document**: These are considered to be semi-structured parts of a Wikipedia article. Every page in Wikipedia consists of a title and subtitles. Considering the data below the subtitles, they can be referred as subparts of the article. These subtitles are partitioned and processed separately in this paper. For example, in the Wikipedia article Rahul Dravid the subtitles can be ‘Early life’, ‘Cricketing Career’, ‘Praises and Accolades’, ‘Teams’, ‘Captaincy Record’, etc.

   **Abstract**: Abstract is the initial few lines of a Wikipedia article which provides the gist of the entire page. Abstract also is a semi-structured part of Wikipedia and can be considered as one of the subtitle of the article without a specific title.

   **Infoboxes**: Infobox is typically a tabular representation of key statistics, which includes all important aspects related to the title of the Wikipedia article. For example, Infobox of Sachin Tendulkar article includes his name, full name, nationality, test debut against, place of birth, etc.

3. APPROACH:

   The approach for identifying NEs is detailed in four steps as follows.

3.1 Clustering of Similar Data:

   The Wikipedia articles are clustered based on the Category links. Consider the English Wikipedia article ‘Rahul Dravid’, the page has a category link ‘Indian Cricketers’. Explore the category link ‘Indian cricketers’ which has around 800 list of links to wikipedia articles that talk about cricketers in India. Repeat the process to explore all category links of given article ‘Rahul Dravid’. This turns out to be 3,853 English Wikipedia articles talking about cricketers. Now, the corresponding Hindi and Marathi articles for each page in the English cluster are fetched using the interlanguage links. Thus, forming clusters separately in Hindi, Marathi for the same category (i.e., Cricketers). Hence, the repetition of clustering process in Hindi/ Marathi is avoided.

3.2 Identification of NEs from Infobox:

   Indian Languages such as Telugu, Tamil, Hindi and Marathi are not only short of resources but also short of data present on web. For 23 GB of English Wikipedia data, the corresponding data in Hindi is around 346MB and Marathi is around 156MB. Hence, there is a need to utilize all possible structural information from Hindi/ Marathi data like infoboxes, subtitles, etc. According to our observation, almost all pages in Hindi/ Marathi has Infobox and almost all entries in Infobox are NEs. Hence, Infobox is justified to be of considerable importance for a study here.

   Infobox can be represented as {Key, Value} pairs, where Key is a common attribute shared across different pages. Every Key is mapped with a Value specific to a Wikipedia article. E.g., a page on ‘Mahender Singh Dhoni’ has Keys such as {Name, Place of Birth, Date of Birth, Year of Birth, debut against, last test against, etc.}, which can be shared across pages. The Values are {Dhoni, Bihar, 7 july, 1981, Sri Lanka, Australia, etc.}, which are specific to a page.

   The Key attributes of an Infobox play a crucial part in identification of NEs. Our observation is supported by the following three facts:

   1. The Key attributes across Wikipedia pages in a cluster (e.g., List of Cricketers) are almost similar.

   2. The Key attributes of Indian language articles (e.g., Hindi, Marathi) are mostly the translated versions of Key attributes of English articles from same cluster, e.g., Name - Nam (meaning Name in Hindi).

   3. In a given cluster, the order of occurrences of the Key attributes are also similar across pages. E.g., Name appears initially, followed by Date of Birth, etc.
The approach of identifying NEs from Infobox is detailed in below two sections.

3.2.1 Map corresponding Keys across Languages:

The Key attributes in an English Wikipedia page can be \{Name, Country, test debut against, etc.\}. Similarly, the Key attributes in a Hindi Wikipedia page can be \{Nam, desh, ... etc\}. Mapping corresponding Keys across languages suggest Name in English should be mapped with Nam (meaning Name) in Hindi.

In order to achieve such mappings, consider a cluster, say 'List of Cricketers'. Extract a page, say 'Sachin Tendulkar', from English and Hindi. Fetch the lists of Key attributes from English and Hindi pages. Map each Key attribute in English with every Key attribute in Hindi. This would result in mappings like, \{Name, Nam, 1, 1\}, \{Name, desh, 1, 0\}, \{Country, Nam, 1, 0\}, \{Country, desh, 1, 1\}, etc. Along with the Key attributes of English and Hindi, each of the above mappings include the number of times they co-occurred and the number of times they occurred in the same order (first attribute of English set of keys and first attribute of Hindi set of keys said to be occurred in same order and likewise). This process is repeated with the remaining pages in the cluster and on every new occurrence of an already existing pair, the corresponding values (+1 for number of co-occurrences and +1 for order of occurrence in case they occurred in same order) are added. Finally, we would have mappings similar to \{Name, Nam, 5, 4\}, \{Name, desh, 4, 0\}, \{Country, Nam, 5, 1\}, \{Country, desh, 5, 4\}, etc.

For each pair, a score is assigned based on a weighted linear combination of the co-occurrence statistics.

\[
\text{Score}(e_i, h_j) = \frac{(\lambda_1 \times \text{number of occurrences}) + (\lambda_2 \times \text{order of occurrence})}{\lambda_1 + \lambda_2 = 1}
\]

The \(\lambda_1\) and \(\lambda_2\) indicate the relative importance assigned to the two different statistics, constrained to Eq.(2). Their values are determined experimentally.

The pair with the highest score is determined as a valid map (e.g., Name - Nam). Similarly, other mappings are identified with the rest of the pairs. This procedure is also repeated for finding mappings across English and Marathi Keys.

3.2.2 Tagging non-English Data with NE tags:

As a result of the previous step, we have obtained mappings of Keys across languages. These Key mappings when applied on a certain Wikipedia article, say 'Mahendra Singh Dhoni', would takes values specific to that page. The Key mappings can then be extended to map their associated Values. E.g., Name - Nam when applied on 'Dhoni' page would result in a Value map of 'Dhoni' - 'महेंद्र सिंह धोनी' (Dhoni in Hindi).

As a preprocessing step, Stanford NER is used to identify NEs and their associated tags on all English articles. With this domain knowledge of list of English NEs and their tags, the Value maps can be tagged similarly. Since, 'Dhoni' would be tagged as 'PERSON' in English, its Hindi counterpart would take up the same NE tag. This process is repeated with all other pages and all Keys to get as many tags possible in the Hindi data. Hence, the Hindi words are tagged without any language expertise.

3.3 Identification of NEs from Abstract:

Abstract briefly summarizes the page, hence it is a good source for NEs. Identification of NE’s from abstract can be explained in three steps: Preprocessing, Word co-occurrences and Tagging of words.

In the preprocessing step, consider abstract of each article from English Wikipedia and tag them using Stanford NER to prune non-NE’s. Then consider abstract of each Hindi/ Marathi article and eliminate the stopwords from it. The stopwords list is generated by considering words that occur above a certain threshold (40%) in the overall dataset.

In the word co-occurrences step, consider the abstract of each article and map each NE tagged English word with every non-tagged Hindi/ Marathi word and assign a default weight (1). The process is repeated with other page abstracts in a cluster. Though, the number of NE’s from English and remaining words from Hindi/ Marathi are relatively low, on every occurrence of a NE in Hindi/ Marathi there will be an occurrence of same NE in English. Hence, if the existing pair of tagged English word and non-tagged Hindi/ Marathi word when occurred in other page abstract, the weight of that pair increases (by 1). Thus, the term co-occurrences between English and other language words are calculated.

Finally in the tagging of words step, Hindi/ Marathi word is mapped with maximum co-occurred (maximum weight) English word. Hence the tag of English word is duplicated to its corresponding Hindi/ Marathi word. Hence Hindi/ Marathi data is tagged.

3.4 Identification of NEs from Subtitles of the Text:

Subtitles divide the content of an article into precise and specific subparts, such that each subpart covers a definite aspect of an article. The study on Hindi/ Marathi Wikipedia lead to an observation that only around 8% of Wikipedia articles contains subtitles and data corresponding to that subtitles. Hence, this presence of less amount of data in non-English languages need to be handled carefully.

Some of the other interesting observations about subtitles includes the following:

1) In a given English cluster (List of Cricketers in this case), there exists almost similar subtitles. For example: Cricketing career, IPL, etc., exists in almost all articles related to cricketers. Some of these subtitles are even specific to the given cluster. As the subtitle ‘Cricketing career’ distinguishes the article to be a part of cluster-Cricketers.

2) The above observation is also applicable to Hindi/ Marathi clusters. But, relatively the data associated with Hindi/ Marathi articles is quite less. Hence, the number of subtitles obtained are just 57% of the number subtitles in English.

3) In a given English Cluster, order of occurrence of subtitles is similar across different articles. The order of occurrence is
similar even for Hindi/ Marathi cluster. Although, there is always a limitation of data which is the major concern to deal with.

The identification of NEs from subtitles can be done in three steps as explained below:

3.4.1 Mapping the Subtitles of Hindi/ Marathi with English subtitles:

From the observations derived, there are two factors of concern. First, is the limitation of data. That is, the number of subtitles in Hindi/ Marathi vary largely from number of subtitles in English. The next major concern is that the mapping need to be language independent which is a challenging task.

Given a cluster 'List of Cricketers', consider each article in English and corresponding article in Hindi/ Marathi then collect subtitles from both the languages. Now map each English subtitle with every Hindi/ Marathi subtitle. However each such mapping is attached with two kinds of weights.

a) The number of occurrences: the number of times the particular pair occurs together throughout the dataset.

b) The order of occurrence:

\[
\text{mod}(pos1 - pos2)/\max(pos1, pos2) \tag{3}
\]

where pos1 is position of English subtitle and pos2 is position of Hindi subtitle in their respective article.

For each pair, a score is assigned based on a weighted linear combination of the co-occurrence statistics.

\[
\text{Score}(e_i, h_j) = (\beta_1 \ast \text{number of occurrences}) + (\beta_2 \ast \text{order of occurrence}) \tag{4}
\]

w.r.t \[\beta_1 + \beta_2 = 1 \tag{5}\]

The values of \(\beta_1\) and \(\beta_2\) given relative importance for the three factors stated above and their values are determined experimentally.

Finally, the Hindi subtitle which has maximum score with English subtitle is mapped. Hence, mapping of subtitles is done by overcoming the above mentioned limitations.

3.4.2 Clustering of Similar Subtitles:

From previous step, we get a list of subtitles mapped in different languages. For each mapping, it consists of English and its corresponding Hindi/ Marathi subtitle. Consider each subtitle (for example: International Cricket career) and the data of it in English. Cluster this data with the corresponding Hindi/ Marathi subtitle data. This cluster relates to one page of Wikipedia dataset. Now repeat this for all pages and for all subtitles. Finally we will get a set of clusters for each subtitle and sets of such subtitles. This step maps similar content of different languages. 

This is one of the important contributions of the paper which has the potential to be applied elsewhere.

3.4.3 Term Co-occurrences:

From previous step, we get a list of subtitles where each subtitle has a set of clusters in English and Hindi/ Marathi. Now, apply the similar approach described in section 3.3. That is, first preprocess the English data by tagging NE’s using Stanford NER. Then, prune Hindi/ Marathi data by eliminating the stopwords from it.

Now for each cluster of a subtitle, there exists a set of tagged English and non-tagged Hindi/ Marathi words. Thereafter, each tagged English word is mapped with every non-tagged Hindi word and assign a default weight (=1) for each such pair. This process is repeated with other clusters within that subtitle. The weight of a pair increases(by 1) for every occurrence of an existing pair.

Finally the Hindi/ Marathi word is mapped with the maximum co-occurred (maximum weightage) English word. As the English word is tagged, this tag is assigned to its Hindi/ Marathi counterpart. The process is repeated for all subtitles and the corresponding Hindi/ Marathi data is tagged. Hence, using the English Stanford NER as an anchor, any non-English data can be tagged.

Some of the concrete examples of NE’s are listed below:

1. Virender Sehwag/PERSON in English has its corresponding NE’s as वीरेंद्र शेखर /PERSON in Hindi and विरेंद्र शेखर /PERSON in Marathi.

2. London/LOCATION NE in English is maximum co-occurred with लंदन/LOCATION in Hindi and लंदन/LOCATION in Marathi.

3. Similarly the NE Indian Premier League/ORGANIZATION is attached with इंडियन प्रीमियर लीग /ORGANIZATION in Hindi and इंडियन प्रीमियर लीग /ORGANIZATION in Marathi.

The whole simple idea behind the approach is to find the term co-occurrences between words of English and Hindi/ Marathi. Since, throughout this process we did not use any of the language specific tools; the approach can be extended to any language.

4. EVALUATION:

4.1 Dataset:

Wikipedia is a free, web-based, collaborative, multilingual encyclopedia. There are 283 language editions available as of now. Wikipedia has both structured (e.g., Infoboxes, Categories, Hyperlinks, InterLanguage links, etc.) and semi-structured (content and organization of the page) information. Hence, the richly linked structure of Wikipedia present across several languages (e.g., English, Hindi, Marathi) are being used to build and enhance many NLP applications including NE identification systems. The dataset that we have used is a cluster of 3,853 English, 2,935 Hindi and 2,622 Marathi Wikipedia articles. This paper mainly concentrated on the list of cricketers of various countries in different languages.
4.2 Baseline System:
We have compared our system with a Hindi NER system developed by LTRC (Language Technologies Research Center) \(^1\), IIIT Hyderabad. They have made use of the Conditional Random Fields (CRF) and was able to achieve an F-Measure of 63\%. Their system is reproduced on our dataset with a 5-fold cross validation using spell variations, pattern of suffixes and POS tagging as the features. We have observed their system as our baseline throughout our experiments. There is no available existing system for Marathi NER. Hence, the Marathi results are not compared but were just reported.

4.3 Evaluation Metrics:
Precision, Recall and F-Measure are the evaluation Metrics and they can be defined as follows:

- **Precision**: \( P = \frac{c}{r} \)
- **Recall**: \( R = \frac{c}{t} \)
- **F-Measure**: \( F = \frac{2 \times P \times R}{P + R} \)

where \( c \) is the number of correctly retrieved (identified) named entities, \( r \) is the total number of named entities retrieved by the system being evaluated (correct plus incorrect) and \( t \) is the total number of named entities in the reference data.

5. EXPERIMENTS AND RESULTS:
The Experiments conducted can be broadly classified as follows:

5.1 Experiment 1: Exploiting the Structure of Wikipedia:
One of the major motivation of the paper is to exploit the complete structure of Wikipedia. Hence in this experiment, first consider the articles of Wikipedia as unstructured pages and term co-occurrences between tagged English and non-tagged Hindi/ Marathi pages are calculated. Then, consider each of the steps mentioned in the approach, i.e. Identification of NE’s from subtitles, abstract, infobox are considered in the same order. Thus, we derived the variations of F-Measure scores by introducing each structural aspect of Wikipedia. Throughout this experiment, the values of \( \lambda_1, \beta_1 \) are 0.65 and 0.7 respectively.

Wiki articles as Unstructured pages: As mentioned above, the data in each English Wikipedia article and the data in its corresponding Hindi/ Marathi article is considered. Then calculate the term co-occurrences (refer to section 3.4.3) between them, which results in the F-Measure scores as follows.

Include Subtitles: Now, consider the semi-structural aspect of Wikipedia, the subtitles. As the results show an improvement by assigning a correct tag to Hindi/ Marathi data. But, the increment in result is quite less because of limitation of data in articles.

Include Abstract: Abstract is also considered to a semi-structural part of Wikipedia. By the inclusion of Abstract concept, the results improved even to a better statistics. But, the limitation of presence of abstract in articles reflects in results.

Include Infobox: The convinient tabular representation of Infobox along with its high availability lead to a promising results as mentioned below.

The tables shown below are for both Hindi and Marathi. But, we couldn’t evaluate Marathi system. So, the results are just put down.

<table>
<thead>
<tr>
<th>Hindi WikiNER</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTRC</td>
<td>64.7</td>
<td>74.6</td>
<td>69.2</td>
</tr>
<tr>
<td>Unstructured Pages</td>
<td>68.4</td>
<td>56.7</td>
<td>62.0</td>
</tr>
<tr>
<td>Include Subtitles</td>
<td>73.5</td>
<td>64.3</td>
<td>68.6</td>
</tr>
<tr>
<td>Include Abstract</td>
<td>80.6</td>
<td>68.9</td>
<td>74.3</td>
</tr>
<tr>
<td>Include Infobox</td>
<td>88.5</td>
<td>73.7</td>
<td>80.42</td>
</tr>
</tbody>
</table>

Table 1: Hindi WikiNER

<table>
<thead>
<tr>
<th>Marathi WikiNER</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured Pages</td>
<td>74.3</td>
<td>65.2</td>
<td>69.4</td>
</tr>
<tr>
<td>Include Subtitles</td>
<td>76.7</td>
<td>67.4</td>
<td>71.5</td>
</tr>
<tr>
<td>Include Abstract</td>
<td>82.5</td>
<td>70.3</td>
<td>75.9</td>
</tr>
<tr>
<td>Include Infobox</td>
<td>89.2</td>
<td>74.6</td>
<td>81.25</td>
</tr>
</tbody>
</table>

Table 2: Marathi WikiNER

5.2 Experiment 2: Variations in Lambda Scores:
Given equation 1:

\[
Score_1(e_i, h_j) = \left( \lambda_1 \times \text{number of co-occurrences} \right) + \left( \lambda_2 \times \text{order of occurrence} \right)
\]

w.r.t \( \lambda_1 + \lambda_2 = 1 \)

Where \( \lambda_1 \) is the weight assigned to the factor num_of_co-occurrences \( \lambda_2 \) is the weight assigned to the factor order_of_occurrence

From equation 2, we assumed that \( \lambda_1 + \lambda_2 = 1 \), \( Score_1(e_i, h_j) \) is the weight assigned to pair of English and Hindi Keys.

The change of values of \( \lambda_1 \) and \( \lambda_2 \) will lead to change of values in weight \( Score_1(e_i, h_j) \). The change of weight \( Score_1(e_i, h_j) \) will lead to the change of Key pairs and their associated values. Hence, reflects in the tags of Hindi/ Marathi data and thus, the F-Measure scores. This describes that change in \( \lambda_1 \) and \( \lambda_2 \) will lead to the changes in F-Measure scores. Therefore, the graph below shows the variations in F-Measure values with the variations in \( \lambda_1 \) values.

The values of \( \lambda_1 \) varies from 0.1 to 0.9.

As shown in Figure: 1, with very high values of \( \lambda_2 \) the F-Measure scores are pretty less. Because, though we observed that order_of_occurrence of keys in Infobox are similar across pages. There is always a limitation about the amount of data present because in this case we refer to a single Hindi/ Marathi article. But as the value of \( \lambda_1 \) increases, there is no restriction as to refer a small amount of data because, the

\(^1\)http://ltrc.iiit.ac.in
num_of_co-occurrences is calculated across all pages. Hence, F-Measure score increases. It is observed that, even in case of considering only num_of_co-occurrences also the results are not promisable, as there is always a chance that one Key attribute of English can be co-occured with many Key attributes of Hindi/ Marathi. Thus the optimum values are: $\lambda_1 = 0.65$ and $\lambda_2 = 0.35$.

5.3 Experiment 3: Varying of beta values:

Given equation 5,

$$\text{Score}_2(e_i, h_j) = (\beta_1 \cdot \text{num_of_co-occurrences}) + (\beta_2 \cdot \text{order_of_occurrence})$$

w.r.t $\beta_1 + \beta_2 = 1$

where $\beta_1$ = weight assigned to num_of_co-occurrences
$\beta_2$ = weight assigned to order_of_occurrence

From equation 6, we assumed that $\beta_1 + \beta_2 = 1$

$\text{Score}_2(e_i, h_j)$ = weight assigned to the pairs of English and Hindi subtitles.

Similar to the previous section, with the change in values of weights $\beta_1, \beta_2$ the weight $\text{Score}_2(e_i, h_j)$ is changed, which will inturn result in change of the assignment of Hindi subtitle to the English subtitle. This will lead to change in F-Measure Scores. Hence, the graphs below, shows the variation of values between $\beta_1$ and F-Measure.

The values of $\beta_1$ and $\beta_2$ varies from 0.1 to 0.9.

6. DISCUSSIONS:

For evaluation, we have made the documents of Hindi/ Marathi to be tagged manually and then compared those tags with the tags produced through this approach. The data in Hindi/ Marathi is much less compared to English. Thus, utilising the data completely has been the major concern of the method. From the first experiment conducted, F-Measure score is less than the baseline when the structure of Wikipedia is not considered. Then, accuracies raised step-by-step by the inclusion of each structural aspect of Wikipedia. Finally, the results are encouraging. From the second and third experiments, we could derive the values of unknown variables for equations 1 and 5.

7. CONCLUSIONS:

This paper identifies and extracts the Named Entities for Indian languages. The major concern is for languages where data is very less compared to English. The approach suggested is very simple but efficient. Each and every structural aspect of Wikipedia is considered seperately and NEs of Hindi/ Marathi are found based on NEs of English. From the Experiments, we can conclude that the accuracies increases with the proper utilisation of Wikipedia structure and its data. Hence, Wikipedia-derived system can be used as a supplement to various applications where language-dependent systems are used. Moreover, the approach suggested can be extended to any language as there are no language-dependant tools used in the paper.

References


R. Grishman. The nyu system for muc-6 or where’s the syntax. In the proceedings of Sixth Message Understanding Conference (MUC-6), Fairfax, Virginia, pages 167–195.


