Language-Independent Named Entity Identification using Wikipedia

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

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

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

Abstract

Recognition of Named Entities (NEs) is a difficult process in Indian languages like Hindi, Telugu, etc., where sufficient gazetteers and annotated corpora are not available compared to English language. This paper details a novel clustering and co-occurrence based approach to map English NEs with their equivalent representations from different languages recognized in a language-independent way. We have substituted the required language specific resources by the richly structured multilingual content of Wikipedia. The approach includes clustering of highly similar Wikipedia articles. Then the NEs in an English article are mapped with other language terms in interlinked articles based on co-occurrence frequencies. The cluster information and the term co-occurrences are considered in extracting the NEs from non-English languages. Hence, the English Wikipedia is used to bootstrap the NEs for other languages. Through this approach, we have availed the structured, semi-structured and multilingual content of the Wikipedia to a massive extent. Experimental results suggest that the proposed approach yields promising results in rates of precision and recall.

1 Introduction

Named entity recognition (NER) is an important subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, etc.

The state-of-art NER systems for English produce near-human performance. However, for non-English languages the state-of-art NER systems perform below par. And for languages that have a lack of resources (e.g., Indian Languages) a NER system with a near-human performance is a distant future.

NER systems so far developed involved linguistic grammar-based techniques as well as statistical models. The grammar-based techniques require linguistic expertise and requires strenuous efforts to build a NER system for every new language. Such techniques can be safely avoided when there is a requirement to build a generic NER system for several languages (e.g., Indian Languages). Statistical NER systems typically require a large amount of manually annotated training data. With the serious lack of such manually annotated training data, the task of high-performance NER system projects as a major challenge for Indian languages.

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. However, the goal here is to identify as many NEs available in Indian languages without using any language-dependent tools or resources.

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,

Proceedings of the First Workshop on Multilingual Modeling, pages 11–17,
Jeju, Republic of Korea, 8-14 July 2012.

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) has been used to build and enhance many NLP applications including NE identification systems. However, the existing approaches that exploit Wikipedia for recognizing NEs concentrates only on the structured parts which results in less recall. Our approach concentrates on exploiting structured and semi-structured parts of Wikipedia and hence yielding better results.

The approach used is simple, efficient, easily reproducible and can be extended to any language as it doesn't use any of the language specific resources.

2 Related Work

Wikipedia has been the subject of a considerable amount of research in recent years including Gabrilovich and Markovitch (2005), Milne et al. (2006), Zesch et al. (2007), Timothy Weale (2006) and Richman and Schone (2008). The most relevant work to this paper are Kazama and Torisawa (2007), Toral and Munoz (2006), Cucerzan (2007), Richman and Schone (2008). More details follow, however it is worth noting that all known prior research is fundamentally monolingual, often developing algorithms that can be adapted to other languages pending availability of the appropriate semantic resources.

Toral and Munoz (2006) used Wikipedia to create lists of NE's. They used the first sentence of Wikipedia articles as likely definitions of the article titles, and used them in attempting to classify the titles as people, locations, organizations, or none. Unlike the method presented in our paper, their algorithm relied on WordNet (or an equivalent resource in another language). The authors noted that their results would need to pass a manual supervision step before being useful for the NER task, and thus did not evaluate their results in the context of a full NER system.

Similarly, Kazama and Torisawa (2007) 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 relies upon a part-of-speech tagger.

Cucerzan (2007), by contrast to the above, used Wikipedia primarily for Named Entity Disambiguation, following the path of Bunescu and Pasca (2006). As in our paper, and unlike the above mentioned works, Cucerzan (2007) made use of the explicit Category information found within Wikipedia. In particular, Category and related list derived data were key pieces of information used to differentiate between various meanings of an ambiguous surface form. Cucerzan (2007) did not make use of the Category information in identifying the class of a given entity. It is to be noted that the NER component was not the focus of their research, and was specific to the English language.

Richman and Schone (2008) emphasized on the use of links between articles of different languages, specifically between English (the largest and best linked Wikipedia) and other languages. The approach uses English Wikipedia structure namely categories and hyperlinks to get NEs and then use language specific tools to derive multilingual NEs.

The following are the majors differences between any of the above approaches to the approach followed in this paper.

- No language resource has been used at any stage of NE identification, unlike the above approaches that used at least one of the language dependent tools like dictionary, POS tagger,etc.
- Our approach utilized several aspects of Wikipedia (e.g., InterLanguage links, Cate-

gories, Sub-titles, Article Text), which has been by far the best exploitation of various structural aspects of Wikipedia.

- Language-independent mapping of multilingual similar content (i.e., the parallel/comparable topics or sentences of different languages) can be used as a reference to any future work. Further details can be found in the Section 4.2.

3 Wikipedia Structure

From Wikipedia, we exploited the following three major units:

Category links: These are the links from an article to 'Category' pages, represented in the form of [[Category:Luzerne County, Pennsylvania]], [[Category:Rivers of Pennsylvania]], etc.

InterLanguage links: Links from an article to a presumably equivalent article in another language. For example, in the English language article 'History of India', one finds a set of links including [[hi:भारतीय इतिहास]]. 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. For example, the article regarding Jimmy Wales has subtitles 'Early life and education', 'Career', etc.

4 Architecture

The system architecture involves 3 main steps and are detailed as follows:

4.1 Related Document Clustering:

Hierarchical clustering outputs a hierarchy, a structure that is more informative than the unstructured set of clusters returned by flat clustering. This paper deals with large amounts of semi-structured data and requires structured clusters as output rather than unstructured clusters. Moreover, specifying the number of clusters beforehand is difficult. Hence, we prefer Hierarchical clustering over Flat clustering in rest of the paper. Bottom-up algorithms can reach a cluster configuration with a better homogeneity than Top-Down clustering. Hence, we prefer bottom-up clustering over top-down clustering.

Within bottom-up clustering there are several similarity measures that can be employed namely single-linkage, complete-linkage, group-average and centroid-measure. This single-link merge criterion is local. Priority is given solely to the area where the two clusters come closest to each other. Other, more distant parts of the cluster and the clusters' overall structure are not taken into account. In complete-link clustering or complete-linkage clustering, the similarity of two clusters is the similarity of their most dissimilar members. In centroid clustering, the similarity of two clusters is defined as the similarity of their centroids. Group-average agglomerative clustering or GAAC evaluates cluster quality based on all similarities between documents, thus avoiding the pitfalls of the single-link and complete-link criteria. Hence, in this paper, we made use of the Group-average agglomerative clustering.

We have considered the English Wikipedia articles which contain InterLanguage links to Hindi articles. The English articles are clustered based on the overlap of terms, i.e., the number of common terms present between articles. The clustering algorithm is detailed as follows:

Initially, consider English Wikipedia data, each article in the dataset is considered as a single document cluster. Now, the distance between two clusters is calculated using

$$\text{SIM-GA}(\omega_i, \omega_j) = \frac{1}{(N_i+N_j)(N_i+N_j-1)} \sum_{d_m \in \omega_i \cup \omega_j} \sum_{d_n \in \omega_i \cup \omega_j, d_m \neq d_n} \vec{d_m} \cdot \vec{d_n}$$

where $\vec{d}$ is the length-normalized vector of document d, $\cdot$ denotes the dot product, and N_i and N_j are the number of documents in ω_i and ω_j, respectively. Using group average agglomerative clustering, the pro-

cess is repeated till we reach a certain threshold (set to 0.2) and thus the hierarchical clusters of English data are formed. In order to cluster documents of other languages, we availed the InterLanguage links and structure of English clusters. The InterLanguage links are used in replicating the cluster structure of English Wikipedia articles across other language articles. Therefore, we avoided the repetition of the clustering step for non-English articles. These different language clusters, being interconnected, are further utilized in our approach.

4.2 Mapping related content within interlinked documents:

As the clustering technique used is hierarchical, the intermediate clustering steps are gathered and are called as subclusters. For example, if two clusters (say Diseases, Hospitals) are merged to form a cluster (say Medicine). Then the Diseases, Hospitals are called subclusters for the Medicine cluster.

We measured the average of cosine similarities between the subtitle lists of the articles in a given cluster. If the average similarity exceeds a threshold (set to 0.72), it would mean the articles in the cluster (e.g., Diseases) all share similar subtitles. Otherwise, we go for a subcluster, until the threshold criteria is met. E.g., any two articles of the cluster Diseases share the common subtitles like Symptoms of Disease, Causes, Precautions, etc. This is illustrated in figure 1. As per our observation, the articles of different languages pertaining to same cluster will have same subtitles but depicted in different languages. The Hindi articles of cluster 'Diseases' share the same subtitles with those in English. This is illustrated in figure 2.

In order to map subtitles across languages, in each cluster, consider the non-English article with maximum number of subtitles and its corresponding English article. A lookup in a bilingual dictionary developed by Rohit et al. (2010) would help in mapping certain subtitles. The rest of the subtitles are mapped based on their order of occurrences. The subtitles are likely to occur at the same order in interlinked articles with high number of subtitles. The dictionary is expanded by adding the

Figure 1: Subtitles of Cancer and Multiple Sclerosis

mapped subtitles obtained from such interlinked articles. This process is repeated with the remaining interlinked articles. Rohit et al. had developed the bilingual dictionary availing Wikipedia titles and abstract information. Hence, their approach is language-independent and doesn't hinder our algorithm from being applied to other languages.

Consider each subtitle of an article in a cluster and collect its subtitle data from that article and from its corresponding interlinked article in Hindi. For example, consider the subtitle 'Causes', collect the subtitle data from an English article (say Cancer) and map it with the subtitle data from the Hindi equivalent page on Cancer. We now have a mapping titled 'Causes - Cancer' for the Cancer articles across languages. Repeat this for all articles and group the mappings of common subtitles. Then, a major group 'Causes' is formed. This group will now have a set of mappings like 'Causes - Cancer', 'Causes - Multiple Sclerosis', etc. Thus the multilingual grouping and mapping is done. 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.

4.3 Term co occurrences model:

Consider a map (e.g., 'Causes - Cancer') which contains both English and Hindi data. Given the fact that the usage of English tools doesn't hurt the extensibility of the approach to other languages, the English data is annotated with Stanford NER and the NEs are retrieved. Hindi data is preprocessed by removing the stop words. The stop words list is generated by considering words that occur above a certain frequency in the overall dataset.

Figure 2: Subtitles of Cancer article across languages

For a given map and preprocessed data, every English NE is paired with every non-tagged Hindi word. Attach a default weight (=1) for each pair. Hence, a pair may look like (tagged English word, non tagged Hindi word, 1). This step is repeated with all other mappings present in a group (Ex: 'Causes - Cancer', 'Causes - Multiple Sclerosis' in the group 'Causes'). On repeated occurrence of the same pair, weight of that pair increases (by 1). Finally, for a English NE term, the Hindi term with which it has highest frequency is identified. Then the NE tag of English term is assigned to Hindi term. Hence, Hindi word is labeled. This step is repeated with the remaining English NEs and Hindi terms.

For example, consider two small mappings, each with two English NEs and one sentence in Hindi. Consider the first map, with "Alexander/PERSON", "India/LOCATION" as English NEs and एलेक्ज़ेंडर ने भारत में पंजाब तक के प्रदेश पर विजय हासिल की थी। as Hindi sentence. Then each NE of English is attached with each Hindi word (except the stop words) like Alexander - एलेक्ज़ेंडर, Alexander - भारत, Alexander - पंजाब, India - एलेक्ज़ेंडर, etc., in all combinations. Consider the second map with 'Alexander/PERSON', 'Philip/PERSON' as English NEs and एलेक्ज़ेंडर के पिता का नाम फिलीप था। as Hindi sentence. The pairs would be Alexander - एलेक्ज़ेंडर, Alexander - फिलीप etc. Hence, the maximum co occurred pair would be Alexander - एलेक्ज़ेंडर (Alexander in Hindi). Then the NE tag of Alexander/PERSON is attached to एलेक्ज़ेंडर/PERSON. Similarly, for the remaining English NEs and Hindi terms, the maximum co-occurred pair is identified and the Hindi term is tagged.

5 Evaluation and Experimental setup:

As our approach requires InterLanguage links, we are only interested in a subset of English and Hindi Wikipedia articles which are interconnected. There are 22,300 articles in English and Hindi Wikipedia that have InterLanguage links. The output of Hierarchical GAAC clustering on this subset was observed to be 345 clusters. We have manually tagged Hindi articles of 50 random clusters (as cluster size can dictate accuracies) with three NE tags (i.e., Person, Organization, Location), resulting in 2,328 Hindi articles with around 11,000 NE tags. All further experiments were performed on this tagged dataset. Precision, Recall and F-measure are the evaluation metrics used to estimate the performance of our system.

In order to compare our system performance with a baseline, we have availed the Hindi NER system developed by Gali et al. (2008) at LTRC (Language Technologies Research Center) [1] that recognizes and annotates Hindi NEs in a given text using Conditional Random Fields (CRF) as the sequential labeling mechanism. 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.

6 Experiments and Results:

The experiments conducted are broadly classified as follows:

Experiment 1: Using the structure of Wikipedia namely Category terms, we can cluster the articles which are having similar category terms. Another approach for clustering is to consider the Wikipedia page as an unstructured page and then cluster the articles based on the similarity of words present in it. We have performed Hierarchical GAAC based clustering for these experiments.

Experiment 2: Different clustering metrics will yield different accuracies for a given data. Here, we will measure which similarity metric is appropriate

[1] http://ltrc.iiit.ac.in

for the dataset under study following a Category information based clustering of articles.

6.1 Experiment 1: Whether to use structure of the Wikipedia page:

No_Category: *Clustering without using the Category information:* As the first experiment, the articles are clustered based on the article text and not using the category terms.

With_Category: *Clustering using the Category information:* In this experiment, the category terms are used for clustering the documents. The F-measure suggests that category terms better capture the semantics of an article when compared to the text of the article. Adding to the fact that category terms suggest a compact representation of an article whereas the text include noisy terms. The compact representation of articles has proved to be crucial by our next set of experiments.

	Precision	Recall	F-measure
NER_LTRC	64.9	50.6	56.81
No_Category	69.8	62.7	66.05
With_Category	73.5	64.3	68.59

Table 1: Experiment to determine the impact of structure based clustering

6.2 Experiment 2: Similarity metrics for Clustering

SLAC: *Single-linkage Agglomerative Clustering:* Single-linkage algorithm would make use of minimum distance between the clusters as similarity metric. One of the drawback for this measure is that if we have even a single document related to two clusters, the clusters are merged. In Wikipedia, we will not have un-related documents, all the documents will be having a certain overlap of terms with each other. Hence, the number of clusters formed are relatively less compared to other two similarity measures. Thus the measures of Precision, Recall and F-measure are quite less.

CLAC: *Complete-linkage Agglomerative Clustering:* Complete-linkage algorithm would make use of maximum distance between the clusters as similarity metric. This results in a preference for compact clusters with small diameters over long. Hence, the accuracies are improved. The drawback is that it causes sensitivity to outliers.

GAAC: *Group Average Agglomerative Clustering:* Group Average is the average between single-linkage metric and complete-linkage metric. Hence, covers the advantages of the both, overcoming the drawbacks of both metrics to some extent. Thus, the accuracies have improved considerably over previous experiments.

	Precision	Recall	F-measure
NER_LTRC	64.9	50.6	56.81
SLAC	67.6	60.3	63.74
CLAC	70.3	61.1	65.38
GAAC	73.5	64.3	68.59

Table 2: Experiment to evaluate similarity metrics

7 Discussions:

From the above results, we have made the following observations. (I) Experiment 1: The Category information of Wikipedia was able to capture the semantics and represent the articles in a compact way resulting in higher accuracies over the article text information. (II) Experiment 2: As each cluster is processed independently while identifying NEs, the compactness and uniformity of the clusters matter in our approach. This is studied by considering different similarity metrics while forming clusters. Finally, from the experiments we conclude that formation of hard clusters matter more for better results of the approach.

8 Conclusions

This paper proposes a method to identify the NEs in Indian languages for which the availability of resources is a major concern. The approach suggested is simple, efficient, easily reproducible and can be extended to any other language as it is developed under a language-independent framework. Wikipedia pages across languages are merged together at subtitle level and then the non-English NEs are identified based on term-term co-occurrence frequencies. The experimental results conclude that the use of Category information has resulted in compact representations and the compactness of the clusters plays a predominant role in determining the accuracies of the system.

References

Daniel M. Bikel and Richard Schwartz and Ralph M. Weischedel 1999. *An Algorithm that Learns What's in a Name*, volume 34. Journal of Machine Learning Research.

Silviu Cucerzan 2007. *Large-scale named entity disambiguation based on Wikipedia data*. In Proc. 2007 Joint Conference on EMNLP and CNLL, pages 708–716.

Evgeniy Gabrilovich and Shaul Markovitch 2007. *Computing semantic relatedness using Wikipedia-based explicit semantic analysis*. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, pages 1606–1611.

Evgeniy Gabrilovich and Shaul Markovitch 2006. *Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge*. proceedings of the 21st national conference on Artificial intelligence - Volume 2, pages 1301–1306.

Evgeniy Gabrilovich and Shaul Markovitch 2005. *Feature generation for text categorization using world knowledge*. In IJCAI05, pages 1048–1053.

Jun'ichi Kazama and Kentaro Torisawa 2007. *Exploiting Wikipedia as External Knowledge for Named Entity Recognition*. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 698–707.

David Milne and Olena Medelyan and Ian H. Witten 2006. *Mining Domain-Specific Thesauri from Wikipedia: A Case Study*. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pages 442–448.

Antonio Toral and Rafael Munoz 2006. *A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia*. In EACL 2006.

Timothy Weale 2006. *Utilizing Wikipedia Categories for Document Classification*. Evaluation, pages 4.

Torsten Zesch and Iryna Gurevych and Max Mühlhäuser 2007. *Analyzing and Accessing Wikipedia as a Lexical Semantic Resource*. Biannual Conference of the Society for Computational Linguistics and Language Technology.

Alexander E. Richman and Patrick Schone 2008. *Mining Wiki Resources for Multilingual Named Entity Recognition*. ACL08.

Razvan Bunescu and Marius Pasca 2006. *Using Encyclopedic Knowledge for Named Entity Disambiguation*. EACL'06.

Karthik Gali and Harshit Surana and Ashwini Vaidya and Praneeth Shishtla and Dipti M Sharma. 2008 *Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition*. IJCNLP'08.

Rohit Bharadwaj G, Niket Tandon and Vasudeva Varma. 2010 *An Iterative approach to extract dictionaries from Wikipedia for under-resourced languages*. ICON'10.