

A Language-Independent Approach to Identify the Named Entities in under-resourced languages and Clustering Multilingual Documents

Kiran Kumar N, Santosh GSK, Vasudeva Varma

International Institute of Information Technology, Hyderabad, India
{kirankumar.n, santosh.gsk}@research.iiit.ac.in, vv@iiit.ac.in

Abstract. This paper presents a language-independent Multilingual Document Clustering (MDC) approach on comparable corpora. Named entities (NEs) such as persons, locations, organizations play a major role in measuring the document similarity. We propose a method to identify these NEs present in under-resourced Indian languages (Hindi and Marathi) using the NEs present in English, which is a high resourced language. The identified NEs are then utilized for the formation of multilingual document clusters using the Bisecting k-means clustering algorithm. We didn't make use of any non-English linguistic tools or resources such as WordNet, Part-Of-Speech tagger, bilingual dictionaries, etc., which makes the proposed approach completely language-independent. Experiments are conducted on a standard dataset provided by FIRE¹ for their 2010 Ad-hoc Cross-Lingual document retrieval task on Indian languages. We have considered English, Hindi and Marathi news datasets for our experiments. The system is evaluated using F-score, Purity and Normalized Mutual Information measures and the results obtained are encouraging.

Keywords

Multilingual Document Clustering, Named entities, Language-independent.

1 Introduction

Multilingual Document Clustering (MDC) is the grouping of text documents, written in different languages, into various clusters, so that the documents that are semantically more related will be in the same cluster. The ever growing content on the web, which is present in different languages has created a need to develop applications to manage huge amount of this varied information. MDC has been shown to be a very useful application which plays a major role in managing such varied information. It has got applications in various streams such as Cross-lingual Information Retrieval (CLIR) [1], where search engine takes query in one language and retrieves results in different languages. Instead of providing results as a single long list, the search engine should display them as list of

¹ Forum for Information Retrieval Evaluation - <http://www.isical.ac.in/~clia/>

clusters, where each cluster contains multilingual documents that are similar in content. It encourages users for cluster based browsing which is very convenient for processing the results.

Named entities (NEs) are the phrases that contain names of persons, organizations, locations, times, and quantities. Extracting and translating such NEs benefits many natural language processing problems such as Cross-lingual Information Retrieval, Cross-lingual Question Answering, Machine Translation and Multilingual Document Clustering. Various tools such as Stanford Part-of-Speech (POS) tagger, WordNet, etc., are available to identify the NEs present in high resourced languages such as English. However, under-resourced languages don't enjoy such facility due to the lack of sufficient tools and resources. To overcome this problem we proposed an approach to identify the NEs present in under-resourced languages (Hindi and Marathi) without using any language dependent tools or resources. The identified NEs are then utilized in the later phase for the formation of multilingual clusters. Fig. 1 gives an overview of the proposed approach.

Bilingual dictionaries, in general, don't cover many NEs. Hence, we used a Wiki dictionary [2] instead of bilingual dictionaries to translate Hindi and Marathi documents into English. The Wiki dictionary covers broader set of NEs and is built availing multilingual Wikipedia titles which are aligned using cross-lingual links. During the translation, in order to match a word in a document with dictionary entries, we require lemmatizers to stem the words to their base forms. But as stated earlier, the support of lemmatizers is limited in under-resourced languages. As an alternative, we used Modified Levenshtein Edit Distance (MLED) [3] metric. It solves the problem of 'relaxed match' between two strings, occurring in their inflected forms. This MLED is used in matching a word in its inflected form with its base form or other inflected forms. The rules are very intuitive and are based on three aspects:

1. Minimum length of the two words.
2. Actual Levenshtein distance between the words.
3. Length of subset string match, starting from first letter.

The rest of the paper is organized as follows: Section 2 talks about the related work. Section 3 describes our proposed approach in detail. Section 4 presents the experiments that support our approach. Finally we conclude our paper and present the future work in Section 5.

2 Related Work

The work proposed in [4] uses Freeling tool, common NE recognizer for English and Spanish to identify the NEs present in both the languages. But, it requires languages involved in the corpora to be of the same linguistic family. Such facility is not available for the Indian languages since they don't belong

to a common linguistic family. Work proposed in [5] performed linguistic analysis such as lemmatization, morphological analysis to recognize the NEs present in the data. They represented each document with keywords and the extracted NEs and performed a Shared Nearest Neighbor (SNN) clustering algorithm for forming final clusters. Friburger *et al.* [6] have created their own Named Entity extraction tool based on a linguistic description with automata. The tool uses finite state transducers, which depends on the grammar of proper names. Authors in [7] have used the aligned English-Italian WordNet predicates present in MultiWordNet [8] for Multilingual named entity recognition. In all the above systems discussed, the authors used language dependent resources or tools to extract the NEs present in the data. Hence, such systems face the problem of extendability of their approaches.

The work proposed in [9] did not make use of any non-English linguistic resources or tools such as WordNet or POS tagger. Instead, they used Wikipedia structure (Category, Interwiki links, etc.) to extract the NEs from the languages. The expectation in this paper is that for any language in which Wikipedia is sufficiently well-developed, a usable set of training data needs to be obtained. Clearly, the Wikipedia coverage of under-resourced languages falls short of this requirement. Hence, we propose a completely language-independent approach to extract the NEs present in under-resourced Indian languages (Hindi and Marathi) by utilizing the NEs present in English (a high resourced language). The detailed description of the proposed approach is given in Section 3.

3 Proposed Approach

In this section, we detail the two phases involved in the proposed approach. Phase-1 involves identification of the NEs present in Hindi and Marathi languages. These NEs are later utilized in Phase-2 for the formation of multilingual clusters.

3.1 Phase-1: NE Identification

As mentioned earlier, NEs such as persons, locations, organizations play a major role in measuring the document similarity. All such NEs present in English documents are identified using the Stanford Named Entity Recognizer²(Stanford NER). As a pre-processing step, all the English documents present in the dataset are processed using MontyLemmatiser³ to obtain the corresponding lemma forms. All the English, Hindi and Marathi text documents are then represented using the classical vector space model [10]. It represents the documents as a vector of keyword based features following the “bag of words” notation having no ordering information. The values in the vector are TFIDF scores. Instead of maintaining a separate stopword list for every language, any word that appears in more

² <http://nlp.stanford.edu/ner/index.shtml>

³ <http://web.media.mit.edu/~hugo/montylingua>

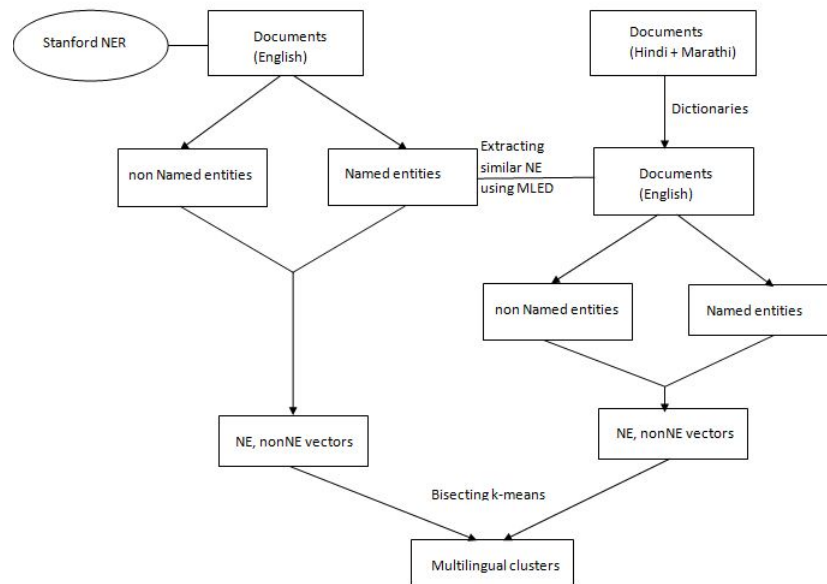


Fig. 1. MDC based on Named entities

than 60% of the documents is considered as a stopword. Hindi and Marathi documents are mapped onto a common ground representation (English) using various dictionaries. Experiments are conducted based on the usage of different dictionaries such as Wiki dictionary [2], bilingual dictionaries (Shabanjali dictionary⁴ and Marathi-Hindi⁵ dictionary). The translated versions of non-English (Hindi, Marathi) documents are also converted into base forms using MontyLemmatiser.

In order to identify the NEs present in non-English documents, the NEs present in all English documents are utilized. All the non-English words after being translated into English are compared with the NEs in English documents and words which have an exact match are identified as the NEs of corresponding non-English documents. After identifying the NEs in all non-English documents, a NE separator function is used to represent each document in the dataset with two vectors namely a NE vector and a nonNE vector. The NE vector contains only NEs present in the document. Whereas, the nonNE vector contains the remaining words of that document. In both these vectors the values are TFIDF scores.

⁴ http://ltrc.iiit.net/onlineServices/ Dictionaries/Dict_Frame.html

⁵ http://ltrc.iiit.net/onlineServices/Dictionaries/Dict_Frame.html

3.2 Phase-2: Multilingual Document Clustering based on Named Entities (MDC_{NE})

Document clustering is an automatic grouping of text documents into clusters so that documents within a cluster have high similarity in comparison to one another, but are dissimilar to documents in other clusters. Various clustering approaches (such as Hierarchical clustering, Partitioning clustering, etc.) are available for cluster formation. Steinbach *et al.* [11] compared different clustering algorithms and concluded that Bisecting k-means performs better than the standard k-means and agglomerative hierarchical clustering. We used Bisecting k-means algorithm for the cluster formation as it combines the strengths of partitional and hierarchical clustering methods by iteratively splitting the biggest cluster using the basic k-means algorithm. As mentioned in the previous section, each document is represented with two vectors namely NE vector and nonNE vector. In the proposed approach (MDC_{NE}), these vectors are linearly combined in measuring the document similarities using Eq. (1) and clustering is performed using Bisecting k-means algorithm. For the evaluation of Bisecting k-means algorithm, we have experimented with fifteen random k values between 30-70 and the average F-score, Purity and NMI values are considered as final clustering results.

The similarity between two documents is calculated by linearly combining the corresponding NE and nonNE vectors. We choose the cosine distance to measure the similarity of two documents (d_i and d_j) which is defined as:

$$sim(d_i, d_j) = \alpha * \left(\frac{dim_1}{dim_h}\right) * sim^{NE} + \beta * \left(\frac{dim_2}{dim_h}\right) * sim^{nonNE} \quad (1)$$

where dim_1 is the dimension of the NE vector, dim_2 is the dimension of the nonNE vector and $dim_h \in \max\{dim_1, dim_2\}$. The sim value is calculated as:

$$sim = \cos(v_i, v_j) = \frac{v_i \cdot v_j}{|v_i| * |v_j|} \quad (2)$$

where vectors v_i, v_j belongs to either NE vector or nonNE vector of documents d_i and d_j respectively. The coefficients α and β indicate the importance of NE vector and nonNE vector respectively ($\alpha + \beta = 1$) in measuring the document similarities. In section 4.1, we present the evaluation criteria to determine the α and β values.

4 Experimental Evaluation

We have conducted experiments using the FIRE 2010 dataset available for the ad-hoc cross lingual document retrieval task. The data consists of news documents collected from 2004 to 2007 for each of the English, Hindi, Bengali and Marathi languages from regional news sources. There are 50 query topics represented in each of these languages. We have considered English, Hindi and

Table 1. Clustering schemes based on different combinations of vectors

Evaluation measure	System-1		System-2		System-3	
	Bilingual dictionaries		Bilingual + Wiki dictionaries		Wiki dictionary	
	MDC_{Keyword} (baseline)	MDC_{NE}	MDC_{Keyword}	MDC_{NE}	MDC_{Keyword}	MDC_{NE}
F-Score	0.553	0.613	0.619	0.660	0.662	0.710
Purity	0.657	0.692	0.687	0.720	0.737	0.762
NMI	0.712	0.743	0.725	0.752	0.761	0.793

Marathi documents for our experiments. We used the topic-annotated documents in English, Hindi and Marathi to build clusters. To introduce noise, we have added topic irrelevant documents that constitute 10 percent of topic documents. Some topics are represented by 8 or 9 documents whereas others are represented by about 50 documents. There are 2182 documents in the resulting collection of which, 650 are in English, 913 are in Hindi and 619 in Marathi. Cluster quality is evaluated by F-score [11], Purity [12] and NMI [13] measures.

F-score combines the information of precision and recall. To compute Purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by total number of documents (N). High purity is easy to achieve when the number of clusters is large - in particular, purity is 1 if each document gets its own cluster. Thus, we cannot use purity to trade off the quality of the clustering against the number of clusters. A measure that allows us to make this tradeoff is Normalized Mutual Information or NMI.

$$NMI(\Omega, C) = \frac{I(\Omega, C)}{[H(\Omega) + H(C)]/2} \quad (3)$$

where I is the Mutual Information and H is entropy of the system.

$$I(\Omega, C) = \sum_k \sum_j \frac{P(\omega_k \cap c_j)}{N} \log \frac{N * P(\omega_k \cap c_j)}{P(\omega_k) * P(c_j)} \quad (4)$$

$$H(\Omega) = - \sum_k P(\omega_k) \log P(\omega_k) \quad (5)$$

where $P(\omega_k)$, $P(c_j)$, and $P(\omega_k \cap c_j)$ are the probabilities of a document being in cluster ω_k , class c_j , and in the intersection of ω_k and c_j , respectively.

The accuracy of the NE identification System of Phase-1 is determined using the following measures:

$$NE_{Precision} = \frac{NE_{s_{correctlyIdentified}}}{NE_{s_{totalNEsIdentified}}} \quad (6)$$

$$NE_{Recall} = \frac{NE_{s_{correctlyIdentified}}}{NE_{s_{totalNEsPresent}}} \quad (7)$$

4.1 Discussion

For the evaluation of NE identification system, we have randomly selected 90 documents from Hindi and Marathi dataset. Three experts from the linguistics department are given 30 documents each to manually identify the NEs present in those documents. The accuracy of NE identification system is determined using Eq. (6) and Eq. (7) and the results obtained are shown in Table-2. For evaluation of the Bisecting k-means algorithm, we have experimented with fifteen random k values between 30-70 and the average F-score, Purity and NMI values are considered as the final clustering results. Three systems are formed based on the usage of different dictionaries such as Wiki dictionary, bilingual dictionaries (Shabanjali dictionary and Marathi-Hindi dictionary) in cluster formation. Linear combinations of NE vectors and nonNE vectors are examined for calculating document similarities in cluster formation.

In System-1, Hindi and Marathi documents are translated to English using only bilingual dictionaries and clustering is performed based on keywords, which is considered as the *baseline* for our experiments. In System-2, both Wiki dictionary and bilingual dictionaries are combinedly used to translate Hindi and Marathi documents into English. Whereas in System-3, only Wiki dictionary is used for the translation. The results obtained in System-1, System-2 and System-3 are shown in Table-1. Results obtained in System-2 shows that clustering based on keywords using bilingual dictionaries and Wiki dictionary together has yielded better results over baseline. Whereas the results obtained in System-3, which is a language-independent approach, shows that clustering based on keywords using only Wiki dictionary has yielded better results over System-2 and the baseline. This might be due to the fact that using bilingual dictionaries create the problem of word sense disambiguation, whereas in Wiki dictionary this problem is evaded as titles of wikipedia are aligned only once. In all the three systems clustering based on NEs has yielded better results over clustering based on keywords which shows the importance of the proposed approach. The reason for improvement in the results might be due to the fact that Wiki dictionary covers broader set of NEs which play a major role in clustering. In all these three systems the α value is determined in training phase, details of which are explained below.

Training Phase Training data constitutes around 60% (1320 documents) of the total documents in the dataset. In these 1320 documents, 400 documents are in English, 550 are in Hindi and 370 in Marathi. The α value is determined by conducting experiments on the training data using Eq. (1). Bisecting k-means algorithm is performed on the training data by varying the α values from 0.0 to 1.0 with 0.1 increment ($\beta = 1-\alpha$). Finally, α is set to the value for which best cluster results are obtained. In our experiments, it is found that setting α value to 0.8 and β to 0.2 has yielded good results in System-1 and System-3. Whereas setting α to 0.7 and β to 0.3 has yielded good results in System-2.

Testing Phase Test data constitutes around 40% (862 documents) of the total

documents in the dataset. Out of these 862 documents, 250 documents are in English, 363 are in Hindi and 249 in Marathi. In all the three systems, Bisecting k-means algorithm is performed on the test data, after setting the α and β values obtained in training phase, using Eq. (1) in similarity calculation.

5 Conclusion and Future work

In this paper we proposed an approach to identify the NEs present in under resourced languages by utilizing the NEs present in English. Bisecting k-means algorithm is performed for clustering multilingual documents based on the identified NEs. The results showcase the effectiveness of the NEs in clustering multilingual documents. From the results it can be concluded that NEs alone are not sufficient for forming better clusters. NEs when combined along with the nonNEs have yielded better clustering results. Our approach is completely language-independent as we haven't used any non-English linguistic resources (such as lemmatizers, NERs, etc.) for processing Hindi and Marathi documents. Instead, we have created alternatives such as Wiki dictionary (built from Wikipedia) and MLED, which is a replacement for lemmatizers. The proposed approach is easy to re-implement and especially useful for the under-resourced languages where the availability of the tools and resources such as dictionaries, lemmatizers, Named Entity Recognizer, etc., is a major problem.

We plan to extend the proposed approach which implements only static clustering to handle the dynamic clustering of multilingual documents. Also, we would like to consider comparable corpora of different languages to study the applicability of our approach.

References

1. Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K.: Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval* **4** (2001) 209–230
2. Kumar, N.K., Santosh, G., Varma, V.: Multilingual document clustering using wikipedia as external knowledge. In: *Proceedings of IRFC*. (2011)
3. Santosh, G., Kumar, N.K., Varma, V.: Ranking multilingual documents using minimal language dependent resources. In: *Proceedings of 12th International Conference on Intelligent Text Processing and Computational Linguistics*, Tokyo, Japan (2011)
4. Montalvo, S., Martínez, R., Casillas, A., Fresno, V.: Multilingual document clustering: an heuristic approach based on cognate named entities. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL)*, Morristown, NJ, USA, Association for Computational Linguistics (2006) 1145–1152
5. Romaric, B.M., Mathieu, B., Besançon, R., Fluhr, C.: Multilingual document clusters discovery. In: *RIAO*. (2004) 1–10

6. Friburger, N., Maurel, D., Giacometti, A.: Textual similarity based on proper names. In: Proceedings of the workshop Mathematical/Formal Methods in Information Retrieval (MFIR2002) at the 25 th ACM SIGIR Conference. (2002) 155–167
7. Negri, M., Magnini, B.: Using wordnet predicates for multilingual named entity recognition. In: Proceedings of The Second Global Wordnet Conference. (2004) 169–174
8. Emanuele Pianta, Luisa Bentivogli, C.G.: Multiwordnet: Developing an aligned multilingual database. In: Proceedings of the 1st International Global WordNet Conference, Mysore, INDIA. (2002)
9. Richman, A.E., Schone, P.: Mining wiki resources for multilingual named entity recognition. (In: Proceedings of ACL-08: HLT)
10. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. *Commun. ACM* **18** (1975) 613–620
11. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: TextMining Workshop, KDD. (2000)
12. Zhao, Y., Karypis, G.: Criterion functions for document clustering: Experiments and analysis. Technical report, Department of Computer Science, University of Minnesota. (2002)
13. Zhong, S., Ghosh, J.: Generative model-based document clustering: a comparative study. *Knowledge and Information Systems* **8** (2005) 374–384