

Effectively Mining Wikipedia for 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 Multilingual Document Clustering (MDC) using Wikipedia on comparable corpora. Particularly, we utilized the cross lingual links, category, outlinks, Infobox information present in Wikipedia to enrich the document representation. We have used Bisecting k-means algorithm for clustering multilingual documents based on the document similarities. Experiments are conducted based on the usage of English and Hindi Wikipedia. We have considered English and Hindi Datasets provided by FIRE'10¹ for Ad-hoc Cross-Lingual document retrieval task on Indian languages. No language specific tools are used, which makes the proposed approach easily extendable for other languages. The system is evaluated using F-score and Purity measures and the results obtained are encouraging.

Keywords

Multilingual Document Clustering, Wikipedia, Document Representation

1 Introduction

MDC is the grouping of text documents written in different languages, into semantically related groups. It plays a significant role in managing huge number of multilingual text documents present in the web. It has got applications in Cross Lingual Information Retrieval (CLIR) [1] systems, training of the parameters in statistical machine translation, among others. The text documents are represented as “bag of words” in traditional text clustering methods. The semantic information of the documents is not considered and this may result in forming false clusters. The most common way to solve this problem is to add semantic information to each document by enriching it’s representation with an external knowledge or an ontology.

In this paper, we focus on using annotated multilingual Wikipedia structure (cross lingual links, outlinks, categories, etc.) in enriching the document representation. The content of a Wikipedia article is annotated by hyperlinks (references) to other articles and they denote the “outlinks” for that article.

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

2 Related Work

MDC is normally applied on parallel [2] or comparable corpus [3]. In the case of the comparable corpora, the documents usually are news articles. In [4], the authors used existing knowledge structure EUROVOC thesaurus for measuring cross lingual document similarity. However the EUROVOC thesaurus supports only European languages. Steinberger *et al.* [5] proposed a method to extract language-independent text features using gazetteers and regular expressions besides thesaurus and classification systems. However, the gazetteers support only a limited set of languages. Hu *et al.* [6] has exploited Wikipedia concepts and categories for monolingual document clustering. In our previous work [7] we have implemented Centroid Similarity based MDC (CS-MDC) using Wikipedia. In this paper we have studied availing Wikipedia in enhancing the performance of MDC based on the Document Similarities (DS-MDC).

3 Proposed Approach

Each document (English or Hindi) in the dataset is represented with a Keyword vector. Three additional vectors namely Category vector, Outlink vector and Infobox vector are obtained by adding semantic information from Wikipedia using Keyword vector. Addition of semantic information to all the terms in the Keyword vector might lead to the distortion of original clustering. Hence, the information is added only for top-n terms based on their TFIDF scores, which also helps in reducing the dimensionality. Various experiments are conducted by varying the top-n value and clustering is performed based on Keyword vector alone. Best clusters are obtained for n=50%. For every term in the Keyword vector, either a Wikipedia article with exact title or a redirected article, if present, is fetched. From this article the outlink, category and Infobox terms are extracted to form the Outlink vector, Category vector and Infobox vector of that document. In all these vectors, the values are the TFIDF scores of those terms. The Keyword vector and the additional vectors are linearly combined for measuring the document similarity.

All the documents are mapped into English using Shabdanjali dictionary² and Wiki dictionary as implemented in [8] availing multilingual Wikipedia titles which are aligned using cross lingual links. We have also implemented Modified Levenshtein Edit Distance proposed in [7] to replace the purpose of Lemmatizers. Two different datasets (Dataset_{HEE} and Dataset_{EHE}) are formed based on the usage of Wikipedia databases (English and Hindi). In Dataset_{HEE}, additional vectors for Hindi documents are obtained from their respective Keyword vectors using Hindi Wikipedia. All these vectors are then mapped into English. In Dataset_{EHE}, Hindi Keyword vectors are initially mapped into English, the additional vectors are then obtained using the English Wikipedia database. In both the datasets additional vectors for English documents are obtained from their respective Keyword vectors using English Wikipedia database. Clustering

² http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Dict_Frame.html

is performed separately on these two datasets based on their document similarities. We used Bisecting k-means algorithm for cluster formation and the results obtained are compared in Table 1. We choose the cosine distance to measure the similarity of two documents (d_i and d_j) which is defined as:

$$sim(d_i, d_j) = sim^{Keyword} + \alpha * sim^{Category} + \beta * sim^{Outlink} + \gamma * sim^{Infobox} \quad (1)$$

Here, $sim(d_i, d_j)$ gives the cosine similarity of the documents d_i, d_j which is calculated as

$$sim = \cos(v_i, v_j) = (v_i \cdot v_j) / (|v_i| * |v_j|) \quad (2)$$

where $v_i, v_j \in \{\text{Keyword, Category, Outlink, Infobox}\}$ vectors of documents d_i and d_j respectively. The coefficients α, β and γ indicate the importance of Category vector, Outlink vector, and Infobox vector respectively.

Table 1. Clustering schemes based on different combinations of vectors

Document Representation	F-Score			Purity		
	CS-MDC	DS-MDC H _E E _E H _H E _E	DS-MDC H _E E _E H _H E _E	CS-MDC	DS-MDC H _E E _E H _H E _E	DS-MDC H _E E _E H _H E _E
keyword (baseline)	0.532		0.606	0.657		0.662
keyword_Category	0.563	0.635	0.625	0.672	0.701	0.683
keyword_Outlink	0.572	0.697	0.651	0.679	0.743	0.705
keyword_Infobox	0.544	0.612	0.609	0.661	0.687	0.673
Category_Outlink	0.351	0.527	0.453	0.434	0.601	0.501
Category_Infobox	0.243	0.420	0.391	0.380	0.507	0.483
Outlink_Infobox	0.248	0.438	0.383	0.405	0.543	0.492
keyword_Category_Outlink	0.567	0.689	0.664	0.683	0.749	0.697
keyword_Outlink_Infobox	0.570	0.653	0.647	0.678	0.726	0.690
keyword_Category_Infobox	0.551	0.622	0.620	0.665	0.699	0.684
Category_Outlink_Infobox	0.312	0.537	0.511	0.443	0.563	0.521
keyword_Category_Outlink_Infobox	0.569	0.681	0.658	0.682	0.739	0.689

4 Experimental Evaluation and Discussion

We have conducted experiments using the English and Hindi documents of FIRE 2010 dataset. There are 50 query topics represented in each of these languages. We used the topic-annotated 1563 documents of which, 650 are in English and 913 in Hindi for our experiments. Cluster quality is evaluated by F-score and Purity measures.

In our experiments, clustering based on Keyword vector is considered as the baseline. Various linear combinations of Keyword, Category, Outlink and Infobox vectors are examined in forming clusters. To determine the α value, experiments are conducted using Equation (1), by varying the α values from 0.0 to 1.0 with 0.1 increment (β and γ are set to 0). The α is set to the value for which best clusters are obtained. Similar experiments are repeated to determine β and γ values. In our experiments, it is found that setting $\alpha = 0.1, \beta = 0.4$ and $\gamma = 0.1$ yielded good results for Dataset $H_E E_E$ whereas $\alpha = 0.2, \beta = 0.1$

and $\gamma = 0.4$ achieved good results for Dataset_H_HE_E. From Table 1, it can be noticed that our experiments using Wikipedia have yielded better results than the baseline in both datasets. As the results obtained by DS-MDC are better than the results obtained by CS-MDC [7], it can be concluded that clustering based on document similarities perform better when compared to clustering based on centroid similarities.

We achieved better clustering results for Dataset_H_EE_E in both the measures when compared to Dataset_H_HE_E. Using only the English Wikipedia database has proved to be beneficial for Dataset_H_EE_E. This might be due the broader coverage of English Wikipedia compared to Hindi Wikipedia. The outlinks information has proved to perform better than categories followed by Infobox information. As the outlinks nearly overlap the context of an article, this might have improved the results better than the rest.

5 Conclusion and Future work

We have performed MDC using Wikipedia. To evaluate the impact of Wikipedia in the cluster formation, we have experimented with English and Hindi Wikipedia databases. Bisecting k-means clustering algorithm is used for the cluster formation and our results showcases the effectiveness of Wikipedia in enhancing MDC performance. We are planning to extend our work by inspecting the role of Named Entities in enriching the document representation for better clustering the multilingual documents.

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. Silva, J., Mexia, J., Coelho, C., Lopes, G.: A statistical approach for multilingual document clustering and topic extraction from clusters. In: *Pliska Studia Mathematica Bulgarica*. (2004) 207–228
3. Romaric, B.M., Mathieu, B., Besançon, R., Fluhr, C.: Multilingual document clusters discovery. In: *RIAO*. (2004) 1–10
4. Steinberger, R., Pouliquen, B., Hagman, J.: Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. In: *CICLing '02*, Springer-Verlag (2002) 415–424
5. Steinberger, R., Pouliquen, B., Ignat, C.: Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. In: *Proc. of the 4th Slovenian Language Technology Conf., Information Society*. (2004)
6. Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: *SIGKDD, ACM* (2009) 389–396
7. Kumar, N.K., Santosh, G., Varma, V.: Multilingual document clustering using wikipedia as external knowledge. In: *IRFC*. (2011)
8. Bharadwaj, G.R., Tandon, N., Varma, V.: An iterative approach to extract dictionaries from wikipedia for under-resourced languages. In: *ICON*. (2010)