www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 4 April 2017, Page No. 20920-20922
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i4.21

A Survey of Multilingual Document Clustering

Mrs.Kavita Moholkar
JSPM's Rajarshi Shahu College of Engineering,
Tathawade, Pune,India
Kavita.moholkar@gmail.com

Abstract: ***The amount of multilingual documents generated on internet, is increasing day by day. Multilingual document clustering (MDC) is a technique of classifying documents in different languages. Classification of documents for the languages without labeled training data set is a major challenge. Two major approaches used till date are machine translation of documents for classification and use bilingual dictionaries for effective translation of trained classification models. This paper surveys various MDC challenge and techniques. The major focus is on the problem of translating documents and classifying it semantically.***

Keywords: Multilingual Text Data; Cross-lingual Text Classification,Clustering**.**

1. Introduction

The massive amount of multilingual documents on the World Wide Web makes the Multilingual document clustering (MDC) problem increasingly important. Typically, MDC refers to the task of classifying documents in different languages using the same taxonomy of predefined categories [1]. The Reuters News Agency, for example, has been using the same taxonomy of subject topics to index International news stories in different languages. Document clustering is defined as classification of text documents into different groups of related documents in an unsupervised manner. Automated classification of multilingual documents is obviously desirable for both cost saving and classification uniformity. However, the documents available are poorly labeled or not labeled at all thus making a MDC more complex.

2. The Multilingual Cluster Identification Problem

Multilingual document clustering approach deals with categorization of multilingual text and cross language information retrieval. Computing the relatedness between multilingual documents is the major challenge for clustering multilingual documens. Clustering is an unsupervised task but the existing multilingual document clustering techniques requires a supervised approach like dictionary, multilingual thesaurus, parallel texts, or comparable corpus etc to achieve cross-lingual semantic interoperability

3. Related Work

The work carried out in the field of Multilingual document clustering (MDC) can be broadly divided into two categories:

1. Machine translation of documents for classification
2. Use bilingual dictionaries

A] Machine Translation Techniques

Bel et al. [9] were amongst the early pioneers examining cross-lingual text categorization. They used the Rocchio algorithm, a popular learning method based on relevance feedback, and the Winnow algorithm, a method for learning a linear classifier from labeled examples, to categorize documents in multiple languages. Roark and Fisher [5] use supervised Machine Learning approach to obtain a query focused sentence ranking. Genetic algorithms is used to carry out the summarization task by Friedman [4]. He carried out the work on English and Hebrew languages. Ling et al. [10] also translated target-language documents (Chinese web pages) to a source language (English), and predicted their labels based the labels of the English documents. Shi et.al. [11] attempted to translate classification models across languages. The model consisted of a bag of weighted terms, where the term weights were the learned model parameters based on labeled data. A bilingual dictionary is used to translate each term in the model to target language. EM algorithm was used to handle ambiguities in translation.

B] Use of bilingual dictionaries

Radev[2] use eight types of summarization algorithms for classification of Chinese and English languages only. Nidhi and V. Gupta [3] had considered sports related documents from the Punjabi News Websites as the corpus. They proposed an Ontology based classification and Hybrid Approach for creating ontology. For linguistic approach, gazetteer list was prepared for Punjabi language The method consists of Pre-processing phase, Feature Extraction phase and Processing Phase. They used new Hybrid approach along with Ontology based classification.

Latifur et.el[4] used a clustering algorithms to build a top-bottom hierarchy based on self organizing tree. WordNet and automatic concept selection algorithm was used to identify correct notion for every node in the linguistic hierarchy. Sandeep Chaware[1] proposed semantic matching approach using Q&A approach for ontology building for Hindi and Marathi languages for inference. He used synset, bi-lingual dictionary, ontology for an entered string for carrying out inference and semantic relatedness. He achieved precise results for precision and recall values for ontology construction and computing inference. Daniil [7] used Wikipedia categories and proposed an unsupervised method for bootstrapping domain ontologies. The method consists of selection of subset of concepts relevant for a Computing and Music domain, splitting up into classes and individuals and identifying the relations between the concepts. The relationship is further classified into subclass of, instance of, part of, and generic related to. The domains of was used to evaluate the method. Saraswathi[8] proposed a system for information retrieval on festival domain for English and Tamil. The authors used ontological tree for inter- language conversion that allows user to query in their native language. Naïve algorithm was used for document search and page ranking algorithm in IR phase. Query disambiguation is done by using language grammar rules and bilingual ontology.

3.1 Limitation and Challenges

The observations from the above methods are as follows:

1. Dictionary does not cover entire vocabulary of that particular language, since new words such as proper names are created almost on a daily basis.
2. Word ambiguity also carries the potential to interrupt a clustering algorithm as generally the algorithm only picks the most frequent sense and ignores the others.
3. A multilingual thesaurus is expensive to build. Existing multilingual thesaurus which is frequently used to for multilingual document clustering is Eurovoc, is available in 22 official languages only. No support for Indian Languages available.
4. Classification of documents by machine translation method using parallel text can be done using term-by-document matrix. It creates a multilingual document space. Parallel texts are available for only a few major languages. It is costly to creat parallel texts and currently only few websites maintain parallel documents in several languages.
5. Moreover, the texts only cover limited domains. Several efforts are made to use them as supervised information souse for documents in different domains have failed to produce satisfactory results.

The table 1 provides a summary of various Multilingual document clustering (MDC) and the associated challenges.

Table 1: Summary of various Multilingual document clustering (MDC) methods

Sr.No	Method	Challenges
1	Dictionary	Word ambiguity, limitation of dictionary
2	Multilingual thesaurus	Word ambiguity, limitation of thesaurus, availability
3	Parallel Texts	Parallel text availability
4	Comparable corpora	Relies on words used in common language
5	Machine translation of documents	Dependency on translator, semantic relatedness, word ambiguity.

.

4. Conclusion

The problem of clustering multilingual documents (MDC) into distinct sets of groups based on their topic similarities. The goal of multilingual document clustering (MDC) is to introduce robust clustering methods which can be applied to documents in various languages. In this paper we have discussed different ways to deal with multilingual documents. It is expected that clustering should be unsupervised .But for multilingual document clustering techniques requires the presence of supervisory information (i.e., dictionary, multilingual thesaurus, parallel texts, or comparable corpus) to achieve cross-lingual semantic interoperability). It is also observed that all the methods discussed above deal with foreign languages and very little effort is made for clustering Indian languages.

References

[1] Ruochen Xu,Yiming Yang, Hanxiao Liu, “Cross-lingual Text Classification via Model Translation with Limited Dictionaries”

[2] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel e Z. Zhu, “MEAD - a platform for multidocument multilingual text summarization.,” Proceedings of LREC 2004, Lisbon, Portugal, 2004.

[3] Nidhi and Gupta, V. “Domain based classification of Punjabi text documents using ontology and hybrid based approach,” in Proc. of 3rd Workshop on South and Southeast Asian Natural Language Processing, SANLP, COLING, Mumbai, 2012, pp. 109-122.

[4] Latifur Khan, Feng Luo and I-ling Yen “Automatic Ontology Derivation from Documents”

[5] B. Roark e S. Fisher, “OGI/OHSU baseline multilingual multi document summarization system,” IEEE International Conference on Microelectronic Systems Education, , USA, 2005.

[6] Sandeep Chaware, Srikantha Rao, “Ontology Supported Inference System for Hindi and Marathi”, 2012 IEEE International Conference on Technology Enhanced Education (ICTEE)Year: 2012 Pages: 1 - 6, DOI: 10.1109/ICTEE.2012.62 08649

[7] Daniil Mirylenka and Andrea Passerini “Bootstrapping Domain Ontologies fromWikipedia: A Uniform Approach” Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

[8] Saraswathi, S., Siddhiqaa, A. M., Kalaimagal, K., and Kalaiyarasi, M. “BiLingual information retrieval system for English and Tamil,” Journal of Computing, vol. 2, pp. 85-89, April 2010.

[9] N. Bel, C. Koster, and M. Villegas. Cross-lingual text categorization. In Research and Advanced Technology for Digital Libraries. Springer Berlin Heidelberg, 2003.

[10] X. Ling, G.-R. Xue, W. Dai, Y. Jiang, Q. Yang, and Y. Yu. Can chinese web pages be classified with english data source? InProceedings of the 17th international conference on World Wide Web, pages 969–978. ACM, 2008.

[11] L. Shi, R. Mihalcea, and M. Tian. Cross language text classi_cation by model translation and semi-supervised learning. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1057–1067. Association for Computational Linguistics, 2010.

[12] Susan Dumais, John Platt, David Heckerman, “Inductive Learning Algorithms and Representations for Text

Categorization”, Microsoft Research One Microsoft Way Redmond, WA 98052.

[13] Alex A. Freitas, “A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery”, Postgraduate Program in Computer Science, Pontificia Universidade Catolica do Parana Rua Imaculada Conceicao, 1155. CuritibaPR. 80215-901. Brazil.

[14] Stanislaw Osinski, “ An algorithm for clustering of web search results “, Masters thesis , Poznan University of Technology, Poland, 2003.

Author Profile

Mrs.Kavita Moholkar received the B.E. degree in Computer Science and Engineering from V.Y.W.S’ College of Engineering ,Amravati in 2002 and M.Tech Degree in Information Technology from Bharti Vidyapeeth College of Engineering , Pune in 2011. She has 15 years of teaching experience. Her expertise lies in data mining, web mining and IR and multilingual document classification. She has guided many UG and PG projects. She also has received research grant from University of Pune.