

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 434 – 440

Procedia
Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

Exploiting Wikipedia API for Hindi-English Cross-Language Information Retrieval

Vijay Kumar Sharma* and Namita Mittal

Malaviya National Institute of Technology, Jaipur, Rajasthan, India

Abstract

The rapidly increasing demographics of the internet population and the abundance of multilingual content on the web increased the communication in multiple languages. Most of the people use their regional languages to express their needs and the language diversity becomes a great barrier. Cross-Language Information Retrieval (CLIR) provides a solution for that language barrier which allows a user to ask a query in the native language and get the relevant documents in the different language. In this paper, we proposed a Wikipedia API based query translation approach. Queries are tokenized and multi-words query terms are created using N-gram technique. Wikipedia title and inter-wiki link features are exploited for query translation. Target language documents are retrieved using vector space retrieval model and BM25 retrieval algorithm. Experiment results shows that the proposed approach achieves better results without exploiting any language resources.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016

Keywords: Cross-Language Information Retrieval; Inter-Wiki Link; Wikipedia API.

1. Introduction

Information Retrieval (IR) is a reasoning process that is used for storing, searching and retrieving the relevant information between a document and the user needs¹¹. Global Internet Usage statistics shows that numbers of web access by the non-English users are tremendously increased. But, all of them are not able to express their queries in English¹. Information retrieval tasks are not restricted to only monolingual but also multilingual. The classical IR normally regards the documents and sentences in other languages as unwanted “noise”¹. The need for handling multiple languages introduce a new area of IR that is CLIR. CLIR deals with user queries in one language and target documents in different language and this becomes a serious issue for world communication. A CLIR approach includes a translation approach followed by mono-lingual information retrieval. There are two types of translation approaches namely query translation and documents translation. A lot of computation time and space is elapsed in document translation approach so query translation approach is preferred⁹. Three query translation approaches are discussed in the State-of-art CLIR i.e. *Dictionary-Based Translation (DT)*, *Corpus-Based Translation (CT)* and *Machine Translation (MT)*. MT and CT approach needs a parallel corpus which is not available for resource-poor

*Corresponding author. Tel.: +91 -9784911021.

E-mail address: sharmavijaykumar55@gmail.com

¹Internet World Stats: <http://www.internetworldstats.com>.

Table 1. Wikipedia Hindi-English Inter-wiki Link Statistics.

Datasets	Available Hindi Article	Available English Inter-Wiki Link Article	Percentage of Availability of Inter-Wiki Link Over The Total Available Article
FIRE 2010	4185	3244	77.51%
FIRE 2011	2460	1849	75.16%

languages like Hindi. It is very cumbersome to create such kind of parallel corpus so DT approach is preferred for fast computation^{2,12}. But, the DT approach has an issue of word translation disambiguation as Bilingual dictionary contains multiple translations for each word. Bilingual dictionary contains very few entries for phrases and has low coverage. The proposed approach provides a solution for these two issues. Wikipedia is an online encyclopedia which is editable by users across the World Wide Web. Wikipedia is very helpful for resource-poor languages like Hindi. The Wikipedia structure and content make it amenable to linguistic research. Each Wikipedia article is associated with the unique title and can provide links to same title articles in different languages called inter-wiki links. The Wikipedia title and inter-wiki link features are utilized in the proposed approach for query translation. Many issues are analyzed during experimentation. A user query contains stop words which mix the noise at the time of searching online Wikipedia. If a user searches for a bi-gram “और मीणा” then Wikipedia API² return a page with the title “किरोड़ी लाल मीणा”. So it is necessary to remove *stop words* before query processing. *N-gram term variation* is also occurred in many Wikipedia article title such as N-gram term “राम विलास पासवान” has the Wikipedia title “रामविलास पासवान”. Many Wikipedia articles inter-wiki links provide wrong target language article. As the title of the target language article provided by inter-wiki link of the Wikipedia article “स्टैनफोर्ड विश्वविद्यालय” is “pacific-12 conference”. Many N-gram terms don’t have any Wikipedia title even unigram also. Our analysis with fire 2010 and 2011 data set concluded that around 75% Hindi Wikipedia articles have English inter-wiki link. Wikipedia inter-wiki link statistics for Hindi-English is shown in Table 1. In the proposed approach, queries are tokenized and multi-word terms are created for phrase identification. These query terms are searched in the titles of online source language Wikipedia using Wikipedia API. The inter-wiki link provides the Wikipedia page in target language so this feature is used for query term translation. Vector space retrieval model is used for target document retrieval. Related work is discussed in Section 2. Proposed approach is discussed in Section 3. Experiment results and discussion are presented in Section 4.

2. Related Work

Pingali *et al.*^{3,4} were experimented with Hindi and Tamil to English language. They used Bilingual dictionary for query translation. OOV terms were transliterated using probabilistic algorithm. Target documents were retrieved using extended Boolean model and Vector based ranking model. Makin *et al.*⁵ were experimented with Hindi document collection. Approximate string matching techniques (LCSR, Jaro-Winkler and Levenstein) were explored to exploit a large number of cognates among Indian languages. They were concluded that bilingual dictionary with cognate matching and transliteration achieved better performance than the bilingual dictionary alone. Sethuramalingam *et al.*⁶ were experimented with FIRE 2008 data. Combinations of dictionaries were used for query translation. Named entities and OOV words were translated using CRF-based named entity recognition tool. Documents were retrieved using Lucene’s OKAPI BM25. Jagathanam *et al.*⁸ were exploited Compressed Word Format (CWF) algorithm for named entity transliteration. Jagarlamudi *et al.*⁷ were prepared a Statistical Machine Translation (SMT) system which trained on aligned parallel sentences and a word alignment table was created. Queries were translated in target language with the use of SMT and transliteration technique. Relevant documents were retrieved using a language modeling based retrieval algorithm. Pattabhi *et al.*¹⁰ were experimented with FIRE 2010 Tamil-English language pair. Named entity terms were extracted from Tamil queries and translate them individually. Bajpai *et al.*¹³ were analyzed the CLIR system for various Indian language and a prototype model was suggested. Queries were translated using any one technique including MT, dictionary based and corpora based. A common problem of word disambiguation was resolved using

²<https://pypi.python.org/pypi/wikipedia/>

WSD technique further Boolean, Vector space and Probabilistic model was used for IR. Adarfe *et al.*¹⁴ were extracted English-Dutch parallel pages from Wikipedia using inter-wiki link. Gaillard *et al.*¹⁵ were experimented English-French CLIR. They segmented the query such that Wikipedia mined bilingual dictionary can translate them. Schonhofen *et al.*¹⁶ were exploited Wikipedia hyper-link structure for query term translation disambiguation. Tyers *et al.*¹⁷ were translated the user queries based on Wikipedia inter-wiki link feature. Bharadwaj *et al.*^{18,19} were utilized cross language links to construct parallel sentences. They also used title, infobox, category, and abstract features to construct Wikipedia based bilingual dictionary. Each query word is disambiguated based on contextual information which is collected using title, redirect title, category, subsection, in-links and out-links features. Erdmann *et al.*²⁰ were increased the Wikipedia based dictionary coverage using Inter-wiki link, redirect page, anchor text and forward/backward link features of Wikipedia. They filter out the incorrect term translation pairs based on backward link feature. Bhagavtula *et al.*²¹ were identified named entities in Indian languages. Highly similar English Wikipedia articles are clustered and tagged by Stanford named entity tagger. Named entities in English article are mapped with other language terms in inter-language linked article based on co-occurrence frequency.

3. Proposed Approach

The proposed approach is divided into three steps. (1) *Pre-processing*, where a query string is tokenized. Stop-words are eliminated and Multi-word terms are created using N-gram technique. (2) *Query translation*, where a query N-gram term is translated using Wikipedia title and inter-wiki link features and (3) *Indexing, Retrieval and Evaluation*. The proposed approach is also depicted in Fig. 1.

3.1 Preprocessing

Query string is tokenized and stop words are eliminated to reduce noise in translation. Multi word terms are created using N-gram for phrase identification. In our approach we use tetra-gram, tri-gram, bi-gram and unigram.

3.2 Query translation

Each N-gram term is further searched in online Wikipedia knowledge base using Wikipedia API and extract the titles of all Wikipedia article in the source language. If any title which has target language inter-wiki link, matches with N-gram term then extract the title of target language article using inter-wiki link. Else Merged N-gram is created by removing white space from N-gram term. Merged N-gram is searched in online Wikipedia knowledge base and extract the titles of all wikipedia article in the source language. If any title which has target language inter-wiki link, matches with Merged N-gram then extract the title of target language article using inter-wiki link. Else extract the titles of all Wikipedia article in the source language using N-gram term. Further, select all the source language title which have more than 80% match with N-gram term. If no title is selected with more than 80% match then select maximal match title. If selected title is not more than one then extract title in the target language using inter-wiki link else extract target language titles for all selected titles and extract maximum frequency words from all target language titles. This procedure is followed for tetra-gram and if we don't get any target language translation then same procedure is followed for tri-gram, bi-gram and unigram. In the case of unigram, a little modification needs to be followed in the last step. If maximal matched title is selected in the case of unigram then there is a possibility of the length of selected title is more than one. So target language translation is extracted based on unigram position in source language title.

3.3 Indexing retrieval and evaluation

Terrier³ search engine is used for indexing, retrieval and evaluation. Many retrieval models are supported by terrier such as Vector space, BM25 etc. Vector space retrieval model and BM25 is used in our experiments.

³<http://terrier.org/>

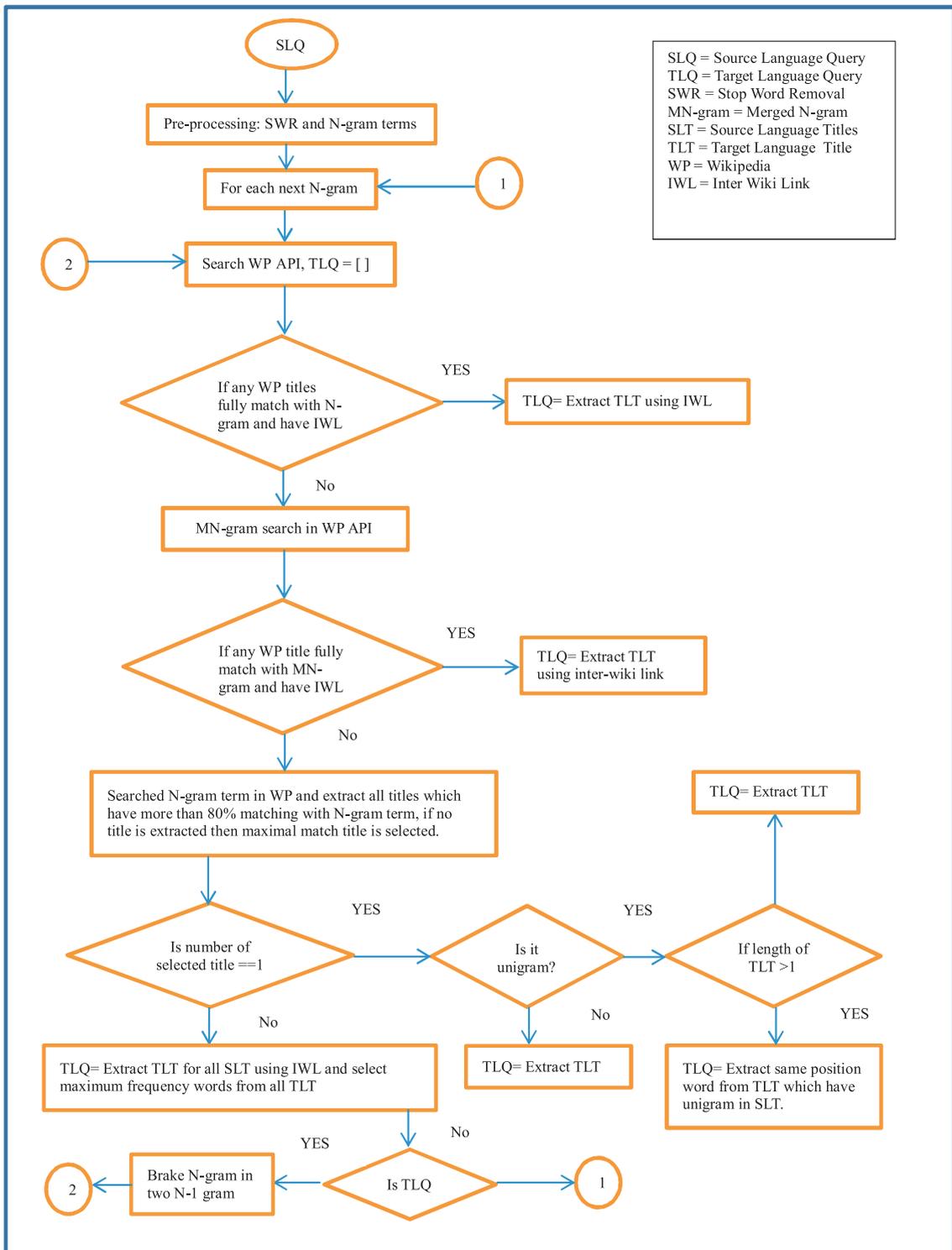


Fig. 1. Proposed Approach for Query Translation.

Table 2. Wikipedia based CLIR Results.

	Experiment Results	Mono Lingual		Cross lingual		
		Fire 2010	Fire 2011	Fire 2010	Fire 2011	
Vector space model	<title> tag	Recall	0.9709	0.8294	0.8315	0.5900
		MAP	0.3705	0.2688	0.1895	0.1096
	<title> and <desc> tag	Recall	0.9954	0.9434	0.9066	0.6914
		MAP	0.4597	0.3584	0.2685	0.1594
BM25 model	<title> tag	Recall	0.9724	0.8290	0.8300	0.5925
		MAP	0.3714	0.2675	0.1899	0.1083
	<title> and <desc> tag	Recall	0.9969	0.9442	0.9081	0.6914
		MAP	0.4650	0.3559	0.2685	0.1601

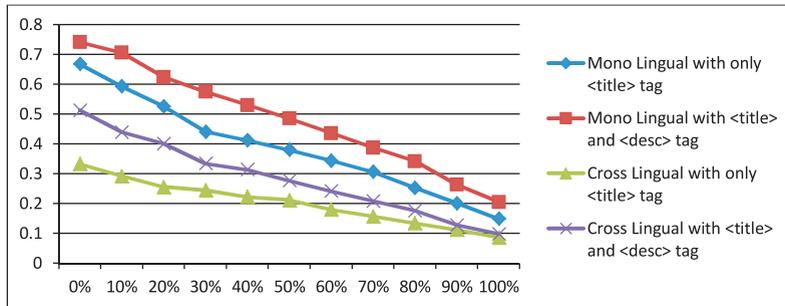


Fig. 2. Precision Score on Recall Points for Fire 2010.

4. Experiment Results and Discussion

The proposed approach is evaluated with FIRE⁴ 2010 and 2011 datasets, which contains a topic set of 50 Hindi language queries and a set of target English language documents. Topic set includes <title>, <desc> and <narr> tag field in each query. We experimented in two ways with both datasets. One is with only <title> tag field and another is with <title> and <desc> tag field. Online Wikipedia Knowledge base is utilized for query translation. N-gram technique is applied to each query and tetra-gram, tri-grams, bi-grams and unigrams are constructed for phrase identification. Terrier search engine is used for indexing, retrieval and evaluation. CLIR system is evaluated by using Recall and Mean Average Precision (MAP). The Recall is the fraction of relevant documents that are retrieved. MAP for a set of queries is the mean of the average precision score of each query. Precision is the fraction of retrieved documents that are relevant to the query. The Proposed approach achieves very good MAP without using any language resource for Hindi-English CLIR as experiment results shown in Table 2.

In the proposed approach, Merged N-gram term is used to eliminate N-gram term variation issue. The retrieval algorithms don't have any significant impact, as the resultant MAP with both the retrieval algorithm is almost same. N-gram terms are used for phrase identification as the term automatically translated based on the context. Wikipedia knowledge base has very poor coverage for the Hindi language, many query terms are unavailable. Around 75% Wikipedia articles have the inter-wiki link and many of them are wrongly linked with target language article. These issues have a very bad impact on MAP. Result analysis of Fire 2010 and Fire 2011 dataset for mono-lingual and cross-lingual retrieval with vector space model presented in Fig. 2 and Fig. 3. The MAP for Fire 2011 is lower than Fire 2010 for both cases of mono-lingual and cross-lingual because fire 2011 topics have short length queries compare to fire 2010. The query terms which do not have any article or inter-wiki link are considered Out of Vocabulary (OOV) terms. These OOV terms may be dictionary words or a named entities.

⁴<http://fire.irsi.res.in/fire/home>

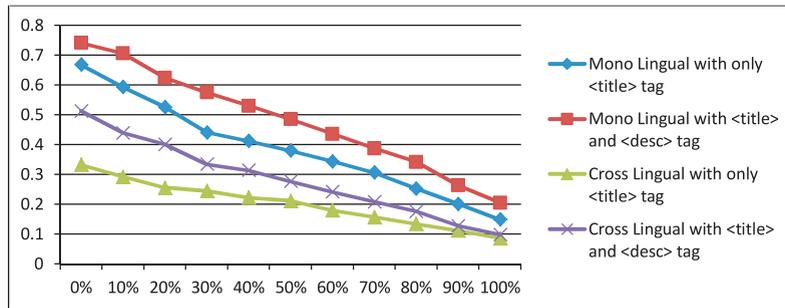


Fig. 3. Precision Score on Recall Points for Fire 2011.

5. Conclusions and Future Work

The Proposed Approach achieved very good MAP without using any language resources. A maximum of 0.2685 MAP achieved with only title and inter-wiki link feature of Wikipedia. Various Wikipedia issues are identified during experimentation. Poor coverage of Hindi Wikipedia article, unavailability of inter-wiki links, wrong target language articles have very bad Impact on MAP. In future, we will improve our system for solving these issues. Wrong target language articles are identified by cross verification of inter-wiki links. Other Wikipedia features such as the article, category, infobox, subsection, hyperlinks will also be utilized in future to solve the OOV term or poor coverage issue of Wikipedia knowledge base.

References

- [1] A. Mustafa, J. Tait and M. Oakes, Literature Review of Cross-Language Information Retrieval, In *Transactions on Engineering, Computing and Technology*, ISSN, (2005).
- [2] V. K. Sharma and N. Mittal, Cross Lingual Information Retrieval (CLIR): Review of Tools, Challenges and Translation Approaches, In *Information System Design and Intelligent Application*, p. 699–708, (2016).
- [3] P. Pingali and V. Varma, Hindi and Telugu to English Cross Language Information Retrieval at CLEF 2006, In *CLEF (Working Notes)*, (2006).
- [4] P. Pingali and V. Varma, IIIT Hyderabad at CLEF 2007-Adhoc Indian Language CLIR Task, In *CLEF (Working Notes)*, (2007).
- [5] R. Makin, N. Pandey, P. Pingali and V. Varma, Approximate String Matching Techniques for Effective CLIR, *International Workshop on Fuzzy Logic and Applications*, Springer-Verlag, pp. 430–437, (2007).
- [6] S. Sethuramalingam and V. Varma, IIIT Hyderabad's CLIR Experiments for FIRE-2008, In *The Working Notes of First Workshop of Forum for Information Retrieval Evaluation (FIRE)*, Kolkata, (2008).
- [7] J. Jagarlamudi and A. Kumaran, Cross-Lingual Information Retrieval System for Indian Languages, In *Advances in Multilingual and Multimodal Information Retrieval*, Springer Berlin Heidelberg, pp. 80–87, (2007).
- [8] S. C. Janarthanam, S. Sethuramalingam and U. Nallasamy, Named Entity Transliteration for Cross-Language Information Retrieval Using Compressed Word Format Mapping Algorithm, In *Proceedings of the 2nd ACM Workshop on Improving non English Web Searching*, ACM, pp. 33–38, (2008).
- [9] N. A. Nasharuddin, M. T. Abdullah, Cross-Lingual Information Retrieval State-of-the-Art, In *Electronic Journal of Computer Science and Information Technology (EJCSIT)*, vol. 2, no. 1, pp. 1–5, (2010).
- [10] R. K. Pattabhi and L. Shobha, AU-KBC FIRE2010 Submission – Cross Lingual Information Retrieval Track: Tamil-English, Fire (2010).
- [11] A. Nagarathinam and S. Saraswathi, State of Art: Cross Lingual Information Retrieval System for Indian Languages, In *International Journal of Computer Application*, vol. 35, no. 13, pp. 15–21, (2011).
- [12] P. Sujatha and P. Dhavachelvan, A Review on the Cross and Multilingual Information Retrieval, In *International Journal of Web & Semantic Technology (IJWesT)*, vol. 2, no. 4, pp. 155–124, (2011).
- [13] P. Bajpai and V. Verma, Cross Language Information Retrieval: In Indian Language Perspective, In *International Journal of Research in Engineering and Technology*, vol. 3, pp. 46–52, (2014).
- [14] S. F. Adarfe and M. D. Rijke, Finding Similar Sentences Across Multiple Languages in Wikipedia, In *The Conference of the European Chapter of the Association for Computational Linguistics*, pp. 62–79, (2006).
- [15] B. Gaillard, M. Boualem and O. Collin, Query Translation Using Wikipedia-Based Resources for Analysis and Disambiguation, In *European Association for Machine Translation*, (2010).
- [16] P. Schonhofen, A. Benczur, I. Biro and K. Csalogany, Cross-Language Retrieval with Wikipedia, In *CLEF*, Springer, pp. 72–79, (2007).
- [17] F. M. Tyers and J. A. Pienaar, Extracting Bilingual Word Pairs from Wikipedia, In *Proceedings of the SALT MIL Workshop at the Language Resources and Evaluation Conference*, LREC (2008).

- [18] R. G. Bharadwaj and V. Varma, Language Independent Identification of Parallel Sentences using Wikipedia, In *Proceedings of the 20th International Conference Companion on World Wide Web*, ACM, pp. 11–12, (2011).
- [19] Bharadwaj, R. G. and V. Varma, Language Independent Context Aware Query Translation using Wikipedia, In *4th Workshop on Building and Using Comparable Corpora*, ACL, pp. 145–150, (2011).
- [20] M. Erdmann and K. Nakayama, Improving the Extraction of Bilingual Terminology from Wikipedia, In *ACM Transaction on Multimedia Computing, Communication and Application*, vol. 5, (2009).
- [21] M. Bhagavatula, S. GSK and V. Varma, Language-Independent Named Entity Identification using Wikipedia, In *Proceeding of the First Workshop on Multilingual Modelling*, pp. 11–17, (2012).