Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 428 – 433

Procedia Computer Science

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

Exploiting Parallel Sentences and Cosine Similarity for Identifying Target Language Translation

Vijay Kumar Sharma* and Namita Mittal

Malaviya National Institute of Technology, Jaipur, Rajasthan, India

Abstract

In recent times, The Internet has become a huge information resource which contains information in multiple languages. Users are not acquainted with all languages and this language diversity becomes a great barrier for world communication. Cross-Language Information Retrieval (CLIR) provides a solution for this language barrier where a user can search the required information in his regional language. In this paper, a CLIR system is proposed based on Parallel Corpus (PC). A set of parallel sentences are extracted from PC which are based on query words. Term frequency matrix and cosine similarity measure are used for identifying target language translation. The proposed Term Frequency Method (TFM) approach is compared with Probabilistic Lexicon Method (PLM) approach and result analysis shows that proposed TFM approach performs better than the PLM approach.

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

Keywords: Cross-Lingual Information Retrieval; Parallel Corpus; Probabilistic Lexicon; Term Frequency.

1. Introduction

Nowadays the internet has overwhelmed by multi-lingual content. The classical IR normally regards the documents and sentences in other languages as unwanted "noise"[1]. Global internet usage statistics shows that the numbers of web access by the non-English users are tremendously increased. But, all of them are not able to express their queries in English[1]. The needs for handling multiple languages introduce a new area of IR that is CLIR. CLIR provides the accessibility of relevant information in a language different than the query language[11]. In CLIR, a user query is translated by either direct translation i.e. *Dictionary-Based Translation* (DT), *Corpus-Based Translation* (CT) and *Machine Translation* (MT) or indirect translation i.e. *Latent Semantic Indexing (LSI), Explicit Semantic Analysis (ESA)* etc.[15]. There are two types of direct 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 preferred[9]. DT approaches have issues of word translation disambiguation and dictionary coverage. MT and CT approach required a parallel corpus. Although it is very difficult to get a parallel corpus but if it is available then CT approach is very effective[2,12]. Most of the researchers were utilized parallel corpus to create

*Corresponding author.

E-mail address: sharmavijaykumar55@gmail.com

[1]Internet World Stats: http://www.internetworldstats.com.

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.092

a probabilistic dictionary. Giza++[2] tool is used to create a probabilistic word alignment table where each word has multiple translations associated with probability score. Query words are translated based on either maximum probability score or Point-wise Mutual Information (PMI) score. Query word translation based on PMI score gives a very poor result because the probability of co-occurrence of two words at sentence level is very low. So in our implementation, we used maximum probability score. GIZA++ training takes much time for creating probabilistic word alignment table. Indirect translation method like LSI uses the parallel corpus to create dual semantic space. LSI method used a relational algebra method, Singular Value Decomposition (SVD) and term-frequency matrix which is very large for the given parallel corpus, so computation cost of LSI method is very high[15]. The proposed approach provides an intermediate solution where a small term frequency matrix is utilized to identify target language translation instead of creating word alignment table by GIZA++, so computation cost is very less. Queries are tokenized and a set of parallel sentences are selected from parallel corpus such that each sentence contains at least one query word. A threshold is empirically defined for selection of parallel sentences for each word to reduce computation cost. Term-frequency matrix is created from selected parallel sentences which contain source language query word vectors and target language sentence word vectors. Cosine similarity measure is used to identify target language translation. Vector space retrieval model is used for target language document retrieval. Related work is discussed in Section 2. The 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 the English language. They used Bilingual dictionary for query translation. Out Of Vocabulary (OOV) terms were transliterated using the 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. Jagarthanam *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 the target language with the use of SMT and transliteration technique. Relevant documents were retrieved using a language modelling 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 analysed 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 WSD technique further Boolean, Vector space and Probabilistic model was used for IR.Pingali *et al.*[14] were used the bilingual lexicon and statistical lexicon created by parallel corpora for query translation. OOV words were transliterated by rule-based method. Mahapatra *et al.*[16] were used GIZA++ tool to get word alignment table from the parallel corpus and sentence word overlap score and WordNet similarity score was used for selecting the best translation. Saravanan *et al.*[17] were created probabilistic translation lexicon by statistical learning on parallel corpora. OOV words were handled with transliteration generation or mining technique. Surya *et al.*[18,19] were used GIZA++ tools to create word alignment table and CRF model was trained on this word alignment table for OOV word transliteration. Larkey *et al.*[20] were used the probabilistic dictionary for query translation. Bradford *et al.*[21] were used Machine Translation software to create parallel corpora and Cross-Lingual LSI method was used for CLIR. Nie *et al.*[22] were created probabilistic lexicon from the parallel corpus and the Probabilistic model combining with bilingual dictionary were used for query translation. Udupa *et al.*[23] were used GIZA++ tool to create a probabilistic lexicon and machine transliteration for OOV words.

[2]http://www.statmt.org/moses/giza/GIZA++.html

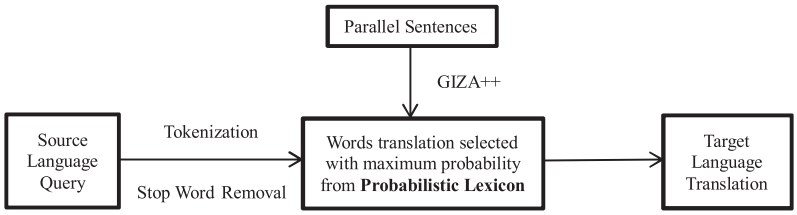


Fig. 1. PLM approach of Query Translation.

3. Proposed Approach

The CLIR approach is divided into two steps. (1) *Query Translation*; where a query string is tokenized, stop-words are eliminated and query words are translated into target language. (2) *Document Retrieval*; where a Vector Space Model (VSM) is used to retrieve target language documents against the translated queries. Two Approaches are proposed for query translation those are Probabilistic Lexicon Method (PLM) and Term Frequency Method (TFM).

3.1 Probabilistic lexicon method

Source language query string is tokenized and stopwords are eliminated to reduce noise in translation. Giza++ tool is used to create a probabilistic lexicon from the parallel corpus. The source language query words translation is selected based on maximum target language translation probability. This method is depicted in Fig. 1.

3.2 Term frequency method

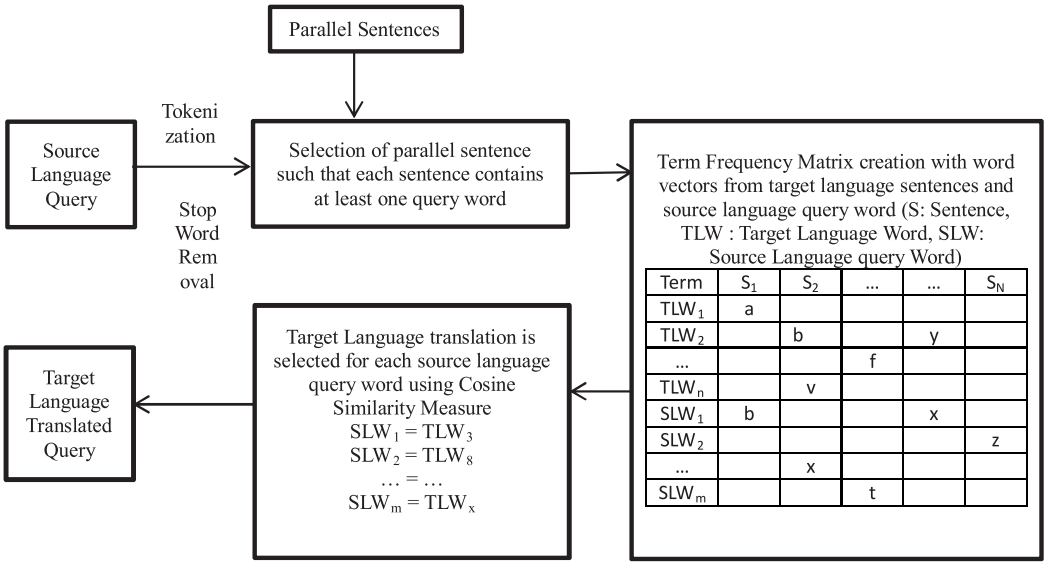
Term frequency method is presented in Fig. 2. Source language query string is tokenized and stop words are eliminated to reduce unnecessary translation. A corpus of parallel sentences are exploited for selection of parallel sentences such that every sentence contains at least one query word. Selected parallel sentences are merged such that each sentence S_i contains source language and target language sentence. A term frequency matrix is created which contains word vectors for terms, where terms are all target language words which are occurred in selected parallel sentences and source language query words as illustrated in Fig. 2. In word vectors, target language word entry with the corresponding sentence will be 1 if target language word is fully matched in a sentence. Source language query word entry with the corresponding sentence will be 1 if source language query word is fully matched in a sentence. If source language query word is not fully matched then extract all source language words from the selected parallel sentences which have the length range between 70% to 130% length of source language query word. Compute longest common subsequence score between source language query word and all the selected words from the selected parallel sentence and if any word get to score more than 75% then the source language query word entry with the corresponding sentence will be 1. Further, Cosine Similarity Score (CSS) is computed for each source language word against all target language words and select target language word with the maximum CSS. CSS computed between two given vectors $A = \{a_1, a_2, \ldots a_N\}$ and $B = \{b_1, b_2, \ldots b_N\}$ is shown in equation 1.

$$CSS = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2}\sqrt{\sum_{i=1}^{N} B_i^2}} \tag{1}$$

4. Experiment Results and Discussion

The proposed approach is evaluated with FIRE[3] 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>

[3]http://fire.irsi.res.in/fire/home



Term	S_1	S_2	...	...	S_N
TLW_1	a				
TLW_2		b		y	
...			f		
TLW_n		v			
SLW_1	b			x	
SLW_2					z
...		x			
SLW_m			t		

Fig. 2. TFM Approach of Query Translation.

Table 1. Comparative Result of PLM and TFM approach of CLIR.

	FIRE 2010		FIRE 2011	
Result	Recall	MAP	Recall	MAP
PLM	0.7488	0.2267	0.6791	0.1672
TFM	0.7519	0.2367	0.6754	0.1623

tag field in each query. We experimented with only <title> tag field. A Hindi-English parallel corpus[44] is exploited in both PLM and TFM approaches. Vector space model is used for indexing and retrieval. CLIR system is evaluated by using Recall and Mean Average Precision (MAP). 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. Comparative result analysis of PLM and TFM is presented in Table 1.

The proposed TFM approach achieves better MAP than the PLM approach. PLM approach takes much time during the training phase and it also requires a huge corpus. The proposed TFM approach does not require huge corpus, it takes only 250 to 500 sentences per query word. So here we get two benefits with TFM approach over PLM approach, i.e. TFM method would be beneficial for the resource-poor language as it does not require huge corpus, and computation time for target language translation is also reduced. The proposed TFM approach is also eliminate the big disadvantage of LSI approach, as LSI approach also used a huge parallel corpus and build a very large matrix which takes a lot of time for computation. The threshold for selecting a number of parallel sentences for each query word is decided empirically which are250 for FIRE 2010 and 500 for FIRE 2011 as shown in Fig 3.

It is very straightforward from the graph that the MAP is approximately equal for every selection of sentences above 90. However Maximum MAP achieved for FIRE 2010 is 0.2637 with 250 sentences. MAP achieved for FIRE 2011 with TFM approach with 500 sentences is approximately equal to PLM approach. Since FIRE 2011 topics set have short queries compare to FIRE 2010. So MAP score is lower and a number of selected parallel sentences are more.

[4]https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-625F-0

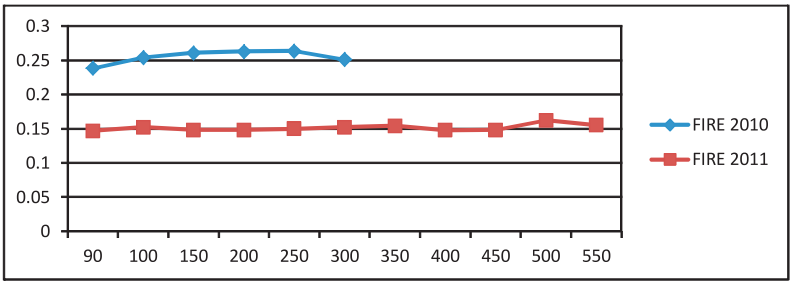


Fig. 3. MAP Score Against Number of Parallel Sentences Selected in TFM Approach.

5. Conclusions and Future Work

The proposed TFM approach achieves very good MAP without using full parallel corpus and also perform fewer computations compare to LSI approach. A maximum of 0.2637 MAP achieved for FIRE 2010 and 0.1623 MAP for FIRE 2011 with only <title> tag. FIRE 2011 topic set have short length queries so their MAP with TFM approach is approximately equal to PLM approach while in the case of FIRE 2010 topic set, MAP achieved with TFM approach is greater than PLM approach. The experiment result analysis shows that the proposed TFM approach is better than PLM approach. In future, the proposed approach will be tested with Wikipedia sentences or parallel sentences extracted from the web.

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] P. Pingali, J. Jagarlamudi and V. Varma, A Dictionary Based Approach with Query Expansion to Cross Language Query Based Multi-Document Summarization: Experiments in Telugu-English, *Mumbai, India*, (2008).
[15] A. Wang, Y. Li and W. Wang, Cross Language Information Retrieval Based on lda, In *International Conference on Intelligent Computing and Intelligent Systems, ICIS 2009, IEEE*, vol. 3, pp. 485–490, (2009).
[16] L. Mahapatra, M. Mohan, M. M. Khapra and P. Bhattacharyya, OWNS: Cross-Lingual Word Sense Disambiguation Using Weighted Overlap Counts and Wordnet Based Similarity Measures, In *Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics*, pp. 138–141, (2010).
[17] K. Saravanan, R. Udupa and A. Kumaran, Crosslingual Information Retrieval System Enhanced with Transliteration Generation and Mining, In *Forum for Information Retrieval Evaluation (FIRE-2010) Workshop*, (2010).

[18] G. Surya, S. Harsha, P. Pingali and V. Verma, Statistical Transliteration for Cross Language Information Retrieval using HMM Alignment Model and CRF, In *Proceedings of the 2nd Workshop on Cross Lingual Information Access*, (2008).
[19] P. Shishtla, G. Surya, S. Sethuramalingam and V. Varma, A Language-Independent Transliteration Schema Using Character Aligned Models at NEWS 2009. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, Association for Computational Linguistics, pp. 40–43, (2009).
[20] L. S. Larkey, M. E. Connell and N. Abduljaleel, Hindi CLIR in Thirty Days, *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 2, no. 2, pp. 130–142, (2003).
[21] R. Bradford and J. Pozniak, Combining Modern Machine Translation Software with LSI for Cross-Lingual Information Processing, In *2014 11th International Conference on Information Technology: New Generations (ITNG)*, IEEE, pp. 65–72, (2014).
[22] J. Nie, M. Simard, P. Isabelle and R. Durand, Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web, In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 74–81, (1999).
[23] R. Udupa, J. Jagarlamudi and K. Saravanan, Microsoft Research India at Fire 2008: Hindi-English Cross-Language Information Retrieval, In *Working Notes for Forum for Information Retrieval Evaluation (FIRE) Workshop*, (2008).