Cross Lingual Information Retrieval (CLIR): Review of Tools, Challenges and Translation Approaches

Vijay Kumar Sharma and Namita Mittal

Abstract Today's Web spreads all over the world and world's communication over the internet leads to globalization and globalization makes it necessary to find information in any language. Since only one language is not recognized by all people across the world. Many people use their regional languages to express their needs and the language diversity becomes a great barrier. Cross Lingual Information Retrieval provides a solution for that barrier which allows a user to ask a query in native language and then to get the document in different language. This paper discusses the CLIR challenges, Query translation techniques and approaches for many Indian and foreign languages and briefly analyses the CLIR tools.

Keywords CLIR · Dictionary translation · Wikipedia translation · UNL · Corpora · Ontology · NER · Google translator · Homonymy · Polysemy

1 Introduction

Information Retrieval (IR) is a reasoning process that is used for storing, searching and retrieving the relevant information between a document and user needs. These tasks are not restricted to only Monolingual but also Multilingual. The documents and sentences in other languages are considered as unwanted "noise" in classical IR [1, 2]. CLIR deals with the situation where a user query and relevant documents are in different language and the language barrier becomes a serious issue for world communication. A CLIR approach includes a translation mechanism followed by mono lingual IR to overcome such language barriers. There are two types of translation namely query translation and documents translation. Query translation

V.K. Sharma (✉) · N. Mittal
Departmemt of Computer Science and Engineering, MNIT, Jaipur, India
e-mail: 2014rcp9541@mnit.ac.in

N. Mittal
e-mail: nmittal.cse@mnit.ac.in

S.C. Satapathy et al. (eds.), *Information Systems Design and Intelligent Applications*, Advances in Intelligent Systems and Computing 433,
DOI 10.1007/978-81-322-2755-7_72

approaches are preferred due to a lot of computation time and space elapsed in document translation approaches [3]. Many workshops and Forums are acquainted to boost research in CLIR. Cross Language Evaluation Forum (CLEF) deals mainly with European languages since 2000. The NII Test Collection for IR System (NTCIR) workshop is planned for enhancing researches in Japanese and other Asian languages. First evaluation exercise by Forum for Information Retrieval Evaluation (FIRE) was completed in 2008 with three Indian languages Hindi, Bengali, Marathi. CLIA consortium includes 11 institutes of India for the project "Development of Cross Lingual Information Access system (CLIA)" funded by government of India. The objective of this project is to create a portal where user queries are responded in three possibilities such as responded in the query language, in Hindi and in English [2]. Literature Survey is discussed in Sect. 2. Issues and Challenges are discussed in Sect. 3. Various CLIR Approaches are discussed in Sect. 4. Section 5 includes Comparative Analysis and Discussion about CLIR translation technique and retrieval strategies. A brief analysis of CLIR tools also included in Sect. 5.

2 Literature Survey

Makin et al. were concluded that bilingual dictionary with cognate matching and transliteration achieves better performance. Parallel corpora and Machine Translation (MT) approaches are not well functioned. [4]. Pirkola et al. were experimented with English and Spanish languages and extract similar terms to develop transliteration rules [5]. Bajpai et al. were developed a prototype model where query was translated using any one technique including MT, dictionary based and corpora based. Word Sense Disambiguation (WSD) technique with Boolean, Vector space and Probabilistic model was used for IR [6]. Chen et al. were experimented with SMT and Parallel corpora for translation [7]. Jagarlamudi et al. were exploited statistical machine translation (SMT) system and transliteration technique for query translation. Language modeling algorithm was used for retrieving the relevant documents [8]. Chinnakotla et al. were used bilingual dictionary and rule based transliteration approach for query translation. Term-Term co-occurrence statistics were used for disambiguation [9]. Gupta et al. were used SMT and transliteration and the queries wise results was undergone mining and a new list of queries was created. Terrier open source[1] search engine was used for information retrieval [10]. Yu et al. were experimented with domain ontology knowledge method which is obtained from user queries and target documents [11]. Monti et al. were developed ontology based CLIR system. First linguistic pre-processing step was applied on source language query then transformation routines (Domain concept

[1] www.terrier.org.

mapping and RDF graph matching) and translation routines (Bilingual dictionary mapping and FSA/FSTs Development) were applied [12].

Chen-Yu et al. were used dictionary based approach and Wikipedia as a live dictionary for Out Of Vocabulary (OOV) terms. Further standard OKAPI BM25 algorithm was used for retrieval [13]. Sorg et al. were used Wikipedia as a knowledge resource for CLIR. Queries and documents both are converted to inter lingual concept space which is either Wikipedia article or categories. A bag-of-concept model was prepared then various vector based retrieval model and term weighting strategies experimented with the conjunction of Cross-Lingual Explicit Semantic Analysis (CL-ESA) [14]. Samantaray et al. were discussed concept based CLIR for agriculture domain. They were used Latent Semantic Analysis (LSA), Explicit Semantic Analysis (ESA) and Universal Networking language (UNL) and WordNet for CLIR and WSD [15]. Xiaoninge et al. were used Google translator due to high performance on named entity translation. Further Chinese character bigram was used as indexing unit, KL-divergence model was used for retrieval and pseudo feedback was used for improve average precision [16]. Zhang et al. were proposed search result based approach and appropriate translation was selected using inverse translation frequency (ITF) method that reduces the impact of the noisy symbols [17]. Pourmahmoud et al. were exploited phrase translation approach with bilingual dictionary and query expansion techniques were used to retrieve documents [18].

3 Issues and Challenges

Various issues and challenges are discussed in Table 1.

Table 1 List of CLIR issues and challenges

Issue and challenges	Homonymy	Polysemy	Word inflection	Phrase translation	Lack of resources	OOV Terms
Definition	Word having two or more different meaning	Word having multiple related meaning	Word may have different grammatical forms	Phrase gives different meaning then the words of phrase	Unavailability of resources for experimentation	Word which not found in dictionary. Like names, new term, technical terms
Example	"Left" means "opposite of right" or "past tense of leave"	"Ring" may be a wedding ring or boxing ring	Good, better, best are different forms of word "Good"	"Couch potato" used for someone who watches too much television	Dictionary, parallel corpora, MT system, character encoding	"H1N1 Malaysia" is a newly added term for influenza disease

Table 2 List of CLIR approaches with description

S. no	Approaches	Description and issues	References
1	Bi-lingual dictionary	Contains a list of source language words with their target language translations. Dictionary quality and coverage is an issue	[1, 2]
2	Corpora based	Corpora are the collection of natural language text in one or multiple languages. Parallel corpora are exactly the translation of each other sentence by sentence or word by word. Comparable corpora are not exactly the translation but cover same topic and contain equivalent vocabulary. Corpora based approach achieves better performance than the bi-lingual dictionary based approach, but these corpora are not available in all languages. In case of unavailability of corpora, it is very cumbersome and computationally expensive to construct parallel corpora of sufficient size	[1, 2]
3	Machine translation (MT)	MT tools used to translate queries into target documents language and target documents into source query language. MT tools save time in case of large text document but short documents are not translated correctly due to lack of context and syntactic structure for WSD. User queries are often short so MT system is not appropriate. MT system is computationally expensive for document translation. MT system is inefficient due to computation cost and unavailability	[20]
4	Transliteration	OOV terms are transliterated by either phonetic mapping or string matching techniques. Phonetic mapping is needed for the languages which have dissimilar alphabets. String matching techniques work best when the two languages having a shared common alphabet. Missing sound is an issue in phonetic mapping. Transliteration variant is an issue in string matching technique	[1, 4]
5	Co-occurrence method	Term-term co-occurrence method is used for translation disambiguation. Only a bilingual dictionary and a monolingual corpus are needed. Monolingual corpus of sufficient size is not available for a large set of languages and it is very cumbersome to create a monolingual corpus	[9, 21]
6	Ontology	An explicit specification for a conceptualization, the combination of ontological knowledge and its connection to the dictionaries gives a powerful approach for resolving CLIR problems	[11, 12]
7	Wikipedia	It is a Web-based, multilingual free content encyclopedia and written by volunteers from the whole world. There are total six million articles in 250 languages and still grow up. Wikipedia inter language link is defined between the same article in different language and it would be useful for translation disambiguation	[13, 14]

(continued)

Table 2 (continued)

S. no	Approaches	Description and issues	References
8	Google translation (GT)	GT is biased towards named entity and Terms in NTCIR topics are mostly name entities thats why Google translation may work well on NTCIR topics	[16]
9	Universal networking language (UNL)	In UNL, a sentence is parsed and a hyper-graph is constructed which having concepts (Universal words) as nodes and relations as arcs. A hyper-graph represents the set of binary relations between any two concepts	[15]
10	Web bases translation	The parallel and comparable web documents are also utilized for query translation and these documents are automatically discovered for different languages. In search result based approach, query terms are disambiguated by search result documents	[17, 22, 23]
11	Word sense disambiguation (WSD)	Appropriate sense of the word is identified. WSD mainly utilize four elements namely first is the word sense selection, second is the external knowledge source utilization, third is the context representation, fourth is the classification method selection	[24]
12	Named entity recognition (NER)	A natural language text is classified into predestined categories such as the person names, locations, organizations etc. State-of-the-art NER systems achieves near-human performance for English language	NER[1]
13	Lemmatization	Every word is simplified to its uninflected form or lemma. For example words “better” and “best” simplified in their uninflected form “good”	[2]

[1] http://en.wikipedia.org/wiki/Named_entity_recognition

4 CLIR Approaches

Various CLIR approaches are discussed in Table 2.

5 Comparative Analysis and Discussion

A comparative analysis of CLIR approaches is presented in Table 3.

Mean Average Precision (MAP) is the evaluation measure. MAP for a set of queries is the mean of the average precision score of each query and precision is the fraction of retrieved documents that are query relevant. Google translator is more

Table 3 Comparative analysis of CLIR approaches

Authors	Languages	Approaches	Datasets	Results (MAP)
Makin et al. [4]	H-TL	BD, CM, TR	BBC Hindi, NavBharat times website	0.2771 (JWS) 0.2449 (LCS)
Jagarlamudi et al. [8]	H-E	MT, PC, TR, LM	CLEF 2007	0.1994 (TD) 0.2156 (TDN)
Chinnakotla et al. [9]	H-E, M-E	BD, TR, COD	CLEF 2007	0.2336 (H (T)) 0.2952 (H (TD)) 0.2163 (M (T))
Chen-Yu et al. [13]	C, J and K	BD, WP, BM25	NTCIR-6	0.0992 (C-CJK-T) 0.0802 (C-CJK-D)
Yu et al. [11]	C-E	BD, HN, OL, COD	NTCIR-4	0.2652 (MITLAB-C-E)
Gupta et al. [10]	H-E	MT, TR, QM, Terrier System	FIRE 2010	0.3723 (BB2C retrieval model)
Sorg et al. [14]	E, G, F, and S	WP, BOC Model, CL-ESA, CAT-ESA, TREE-ESA	JRC-acquis (J) and Multext (M)	0.33 (M), 0.28 (J) (CLESA), 0.43 (M), 0.33 (J) (Cat-ESA), 0.46(M) and 0.31 (J) (Tree-ESA)
Xiaoning et al. [16]	C-E	GT, CCB, KL-Divergence and PF	NTCIR-7	0.3889
Chen et al. [7]	E, G, F, DT, I, S	L and H MT System, PC	CLEF 2003	0.3814 (F-G), 0.3446 (F-DT), 0.3859 (G-I), 0.4340 (I-S), 0.4694 (E-G), 0.4303 (E-S)
Zhang et al. [17]	E-C	SRWB and ITF	NTCIR-4	0.1582
Pourmahm-oud et al. [18]	P-E	BD, CT, QE, LM	Test collection prepared by themselves	0.3648 (without QE) 0.4337 (with QE)

BD bilingual dictionary, *CM* cognate matching, *TR* transliteration, *HN* HowNet, *PC* parallel corpora, *MT* machine translation, *LM* language modelling, *WP* wikipedia, *COD* co-occurrence distance, *OL* ontology, *QM* query mining, *QE* query expansion, *LSI* latent semantic indexing, *BOC* bag of concept, *GT* google translator, *CCB* chinese character bigram, *PF* pseudo feedback, *SRWB* search result web based approach, *ITF* inverse translation frequency, *CT* cohesion translation, *BT* back translation, *ER* entity recognition, *JWS* jaro winkler similarity, *LCS* longest common subsequence, *T* title, *D* description, *N* narration, foreign language (*E* English, *G* Germen, *DT*: Dutch, *I* Italian, *S* Spanish, *C* Chinese, *P* Persian, *F* Finnish, *J* Japanese, *K* Korean, *FR* French), Indian languages (*H* Hindi, *M* Marathi, *TL* Telugu)

effective due to biasing towards named entities and 0.3889 MAP achieved for English-Chinese [16]. Machine translation and Parallel corpora combinedly achieve better MAP that is 0.4694 for English-Germen [7] but lack of resources problem is there because a parallel corpora of enough size is not available for all languages.

Table 4 Comparative analysis of CLIR tools

S. no	Tools	Language supported	Translation technique	Functionality	Limitation
1	MULINEX	F, G and E	BD and BT	Interactive QD and QE, summaries and search results are translated on demand	Synonymy and Homonymy, User assisted query translation
2	KEIZAI	E, J and K	BD and PC	Interactive query translation along with English definition, target documents summary with English summary & document thumbnails visualization	Synonymy and homonymy, User assisted query translation
3	UCLIR	Arabic languages	BD and MT	Multi lingual query, interactive and non-interactive English query, Relevant retrieved document translated in English by word level (dictionary) or document level (MT), document thumbnails visualization	Non-interactive query approach include irrelevant translation, Interactive query approach is user assisted query translation
4	MIRACLE	English and other languages	BD	user can select or deselect some translation, query reformulation, automatic and user assisted query translation	Resources are not available, Homonymy and Synonymy
5	MULTILEX EXPLORER	Support multi lingual	WordNet and Web Search Engine	Exploring context of query, WSD, language selection, QE, automatic categorization, circle visualization	WordNet not available for all languages
6	MULTI SEARCHER	Support multi lingual	BD, PC, ER, Mutual information	User assisted disambiguation, Automatic translation disambiguation deal with the user's lack of knowledge in target language, Automatic Document categorization	Parallel Corpora not available for all languages

Mostly researcher used bilingual dictionary because it is available for all languages and also takes nominal computation cost. Bi-lingual dictionary with Cohesion translation and Query expansion achieves 0.4337 for Persian-English [18]. Wikipedia is used to identify OOV terms but Wikipedia with sufficient data is available for a limited number of languages. CLIR with Wikipedia achieves 0.46 MAP [14].

Ontology, WordNet, UNL and co-occurrence translation used for resolving term homonymy and polysemy issues. Dictionary coverage and quality, phrase translation, Homonymy, Polysemy and Lack of resources are major challenges for CLIR. Many comprehensive tools are cultivated to resolve the language barrier issue, such as MT tools and CLIR tools [19]. A brief study to the CLIR tools is summarized in the Table 4. All these tools uses bilingual dictionary because of nominal time computation. A common problem of user assisted query translation was tried to remove in MIRACLE, MULTI LEX EXPLORER and MULTI SEARCHER. Automatic query translation suffered by a problem of homonymy and polysemy.

6 Conclusion

CLIR enables searching documents via eternal diversity of languages across the world. It removes the linguistic gap and allows a user to submit a query in a language different than the target documents. A CLIR method includes a translation mechanism followed by monolingual retrieval. It is analyzed that query translation always efficient choice than document translation. In this paper, various CLIR issues and challenges and Query translation approaches with disambiguation are discussed. A comparative analysis of CLIR approaches is presented in Table 3. A CLIR approach with Bi-Lingual dictionary, Cohesion Translation, query expansion and Language Modeling achieves good MAP i.e. 0.4337. Another CLIR approach with Wikipedia, Bag of Concept and Cross language- Explicit Semantic analysis achieves better MAP i.e. 0.46. MT with parallel corpora CLIR approach achieves 0.4694 MAP. A brief analysis of CLIR tools is represented in Table 4. Dictionary Coverage and Quality, Unavailability of Parallel Corpora, Phrase Translation, Homonymy and Polysemy are concluded as major issues.

References

1. Nagarathinam A., Saraswathi S.: State of art: Cross Lingual Information Retrieval System for Indian Languages. In International Journal of computer application, Vol. 35, No. 13, pp. 15–21 (2006).
2. Nasharuddin N., Abdullah M.: 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).

3. Oard, D.W.: A Comparative Study of Query and Document Translation for Cross-language Information Retrieval. In Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup., Springer-Verlag, pp. 472–483 (1998).
4. Makin R., Pandey N., Pingali P., Varma V.: Approximate String Matching Techniques for Effective CLIR. In International Workshop on Fuzzy Logic and Applications, Italy, Springer-Verlag, pp. 430–437 (2007).
5. Pirkola A., Toivonen J., Keskustalo H., Visala K., Jarvelin K.: Fuzzy translation of cross-lingual spelling variants. In: Proceedings of SIGIR'03, pp. 345–352 (2003).
6. Bajpai P., Verma P.: Cross Language Information Retrieval: In Indian Language Perspective. International Journal of Research in Engineering and Technology, Vol. 3, pp. 46–52 (2014).
7. Chen A., Gey F.C.: Combining Query Translation and Document Translation in Cross-Language Retrieval. In Comparative Evaluation of Multilingual Information Access Systems, Springer Berlin: Heidelberg, pp. 108–121 (2004).
8. Jagarlamudi J., Kumaran A.: Cross-Lingual Information Retrieval System for Indian Languages. In Advances in multilingual and multi modal information retrieval, pp. 80–87 (2008).
9. Chinnakotal M., Ranadive S., Dhamani O.P., Bhattacharyya P.: Hindi to English and Marathi to English Cross Language Information Retrieval Evaluation. In Advances in Multilingual and Multimodal Information Retrieval, springer-verlag, pp. 111–118 (2008).
10. Gupta S. Kumar, Sinha A., Jain M.: Cross Lingual Information Retrieval with SMT and Query Mining. In Advanced Computing: An International Journal (ACIJ), Vol.2, No.5, pp. 33–39 (2011).
11. Yu F., Zheng D., Zhao T., Li S., Yu H.: Chinese-English Cross-Lingual Information Retrieval based on Domain Ontology Knowledge. In International conference on Computational Intelligence and Security, Vol. 2, pp. 1460–1463 (2006).
12. Monti J., Monteleone M.: Natural Language Processing and Big Data An Ontology-Based Approach for Cross-Lingual Information Retrieval. In International Conference on Social Computing, pp. 725–731 (2013).
13. Chen-Yu S., Tien-Chien L., Shih-Hung W.: Using Wikipedia to Translate OOV Terms on MLIR. In Proceedings of NTCIR-6 Workshop Meeting, Tokyo, Japan, pp. 109–115 (2007).
14. Sorg P., Cimiano P.: Exploiting Wikipedia for Cross-Lingual and Multi-Lingual Information Retrieval. Elsevier, pp. 26–45 (2012).
15. Samantaray S. D.: An Intelligent Concept based Search Engine with Cross Linguility support. In 7th International Conference on Industrial Electronics and Applications, Singapore, pp-1441–1446 (2012).
16. Xiaoning H., Peidong W., Haoliang Q., Muyun Y., Guohua L., Yong X.: Using Google Translation in Cross-Lingual Information Retrieval, In Proceedings of NTCIR-7 Workshop Meeting, Tokyo, Japan, pp. 159–161 (2008).
17. Zhang J., Sun L. and Min J.: Using the Web Corpus to Translate the Queries in Cross-Lingual Information Retrieval. In Proceeding of NLP_KE, pp. 493–498 (2005).
18. Pourmahmoud S., Shamsfard M.: Semantic Cross-Lingual Information Retrieval. In International symposium on computer and information sciences, pp. 1–4 (2008).
19. Ahmed F., Nurnberger A.: Literature review of interactive cross language information retrieval tools. In The international Arab Journal of Information Technology, Vol. 9, No. 5, pp. 479–486 (2012).
20. Boretz, A., AppTek Launches Hybrid Machine Translation Software, in Speech Tag Online Magazine (2009).
21. Yuan, S., Yu S.: A new method for cross-language information retrieval by summing weights of graphs. In Fourth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE Computer Society, pp. 326–330 (2007).

22. Nie, J., Simard M., Isabelle P., Durand R.: Cross-Language Information Retreval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In Proc. OfACM-SIGIR, pp. 74–81 (1999).
23. Lu W., Chien L., Lee H.: Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach. ACM Transactions on Information Systems 22(2), pp. 242–269 (2004).
24. Navigly R.: Word Sense Disambiguation: A Survey. ACM computing survey, Vol. 41, No. 2 (2009).