Dictionary Based Translation Approaches in Cross Language Information Retrieval: State of the Art

B.N.V Narasimha Raju, M S V S Bhadri Raju

Abstract—Information Retrieval (IR) is the process of finding set of documents or texts that are required by the user. In the past decade internet content in English is shrunk from 39 percent to 27 percent. By the end of 2011, the content available in web was 24 percent of Chinese and 27 percent of English. Soon, the Chinese content will overtake English content. Other languages like Japanese, Korean, Hindi, Tamil, Malay, Thai, Vietnamese, Arabic etc. has a gradual growth in the web content. This causes more importance for multilingual approach. So, the user may have the necessity to retrieve the information in another language; this kind of problem is solved by using Cross Language Information Retrieval (CLIR), a sub field of IR. CLIR retrieves the information that is different from user query language. For retrieving information, CLIR presents different methods such as dictionary based translation, Machine translation and Corpus based translation. In these methods the importance of dictionary based translation has been increased due to growth in the availability of machine readable dictionaries. This article presents a detailed review of dictionary based translation, with emphasis on recent developments.

Index Terms— Dictionary based translation, translation ambiguity, Phrases and compound words, Out-of-vocabulary, Transliteration, Bilingual dictionaries, Information retrieval.

1 INTRODUCTION

INFORMSTION Retrieval (IR) is to obtain the relevant information from a collection of information resources. In information retrieval process user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In Information Retrieval a query matches several objects instead of matching a single object uniquely. The top ranking objects are then shown to the user. This ranking of objects in most of the IR system depends on numeric score computed by the matching process of system. In the information era searching is a part of our daily life. Ideally, we show interest to search information in our native language but sometimes the information may not be available. In such a situation we may show interest to search the information in other languages, this arises the problem of cross-language information retrieval. The goal of CLIR is to find the information written in language other than that of query language. The key problem in CLIR is translation, where one should translate the language of query or documents to another language. For this specific translation methods are required. The architecture of CLIR [1] is shown in Fig. 1. In general, CLIR system works on a specific document collection. There are three ways to use the translation module. The first way is to use document translation approach i.e. mapping of the document representation into the query representation space. The second way is to use query translation approach i.e. mapping of the query representation into the document representation space. The third way is to map both document and query representations to a third space. On both, the query and document indexing is carried out to some extent. Relevance feedback on each document describes how strongly the document and query are related. Optionally, once a list of documents is identified by the system a feedback process can take place. Translation techniques in CLIR are categorized into three types. They are dictionary based translation, Machine translation and Corpus based translation. In Dictionary based translation by using a bilingual dictionary, each word or phrase is translated from source language to target language. Dictionaries are ordered according to different principles. In CLIR, dictionaries usually contain a word list and their translations. For a long time, Machine Translation and construction of various translation systems and resources are done manually but without any manual intervention Machine Translation (MT) systems provides full-text machine translation. The quality of translation is also improved than in earlier

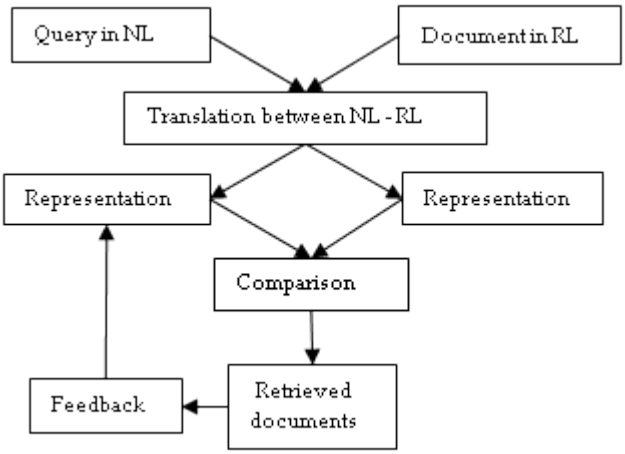

Fig. 1. Architecture of CLIR System.

- *B.N.V Narasimha Raju, Department of Computer Science and Engineering, S.R.K.R Engineering College, Affiliated to Andhra University, Chinna Amiram, Andhra Pradesh 534204, India, PH- +919963322237; E-mail: buddaraju.narasimharaju@gmail.com*
- *M S V S Bhadri Raju, Department of Computer Science and Engineering, S.R.K.R Engineering College, Affiliated to Andhra University, Chinna Amiram, Andhra Pradesh 534204, India E-mail: msramaraju@gmail.com*

times. In corpus based translation knowledge is derived from parallel and comparable corpora. A parallel text is a document written in one language and presented next to its translation in another. Large collections of parallel texts are referred to as parallel corpora. A comparable corpus is a combination of texts that are composed independently but share the same communicative function and theme.

2 Dictionary Based Translation

Dictionary-based approach is getting prominence due to the increasing availability of machine readable dictionaries. Rather than using parallel corpus based methods it is easy to use dictionary-based methods for query translation. In this approach the queries are translated by using bilingual dictionaries in which we look up the some or all of the translated terms. Dictionaries in CLIR contain the word list and their translation. Dictionaries may contain additional information such as definitions and examples that may help in particular context for selection of more appropriate translations. In dictionary based translation the basic approach is to translate word-by-word. In this for a query word we have to select more suitable translation that can be achieved in two ways. Firstly, for a query word we have to use all the translations that may add the inappropriate translations and retrieves the irrelevant documents. This can also be solved by normalization of term weights for translations per source term. The other way is to select first translation in the list that also causes problem as in many dictionaries this assumption is false. These two strategies have low effectiveness in retrieval. The performance of dictionary-based approaches basically depends on three factors: phrases and compound words, translation ambiguity, and out-of-vocabulary terms.

Phrases of a sentence can be identified by using IR parts-of-speech (POS). Identifying and translating the phrases can reduce the impact on phases and compound words. When a query word is being translated then it may cause translation ambiguity i.e. the query word may contain more than one translation and that causes the ambiguity for the selection of appropriate word for translation in that context. When translating a query word that may cause a problem of query term or source term missing in the bilingual dictionary, this type of problem is called as out-of-vocabulary problem.

A method [2] is proposed to combine two bilingual dictionaries which yields the third, using one pivot language. In this case we combine a Japanese-English dictionary with a Malay-English dictionary, to produce a Japanese-Malay dictionary. Normalization of the pivot language has improved the matching of this method. In [3] suggested that many operational IR indexes are non-normalized. As the most translations are lemmas it becomes a challenge for dictionary-based cross-language retrieval. A non-normalized index is a challenge of dictionary-based CLIR. There are two optional approaches for testing. They are Frequent Case Generation (FCG) and s-gramming. For a given lemma we can automatically generate the most frequent inflected forms using FCG. S-gramming is an approximate string matching technique. The language pairs in our tests were English–Finnish, English–Swedish, Swedish–Finnish and Finnish–Swedish. Depending on the language pair, both approaches performed well.

3 Identifying And Translating Phrases And Compound Words

Identifying and translating phrases are used in the query translation to improve the retrieval performance. New compounds words are generated by Natural languages because they are productive systems. Some languages of translation dictionaries include the lexicalized compounds. Most compounds are untranslatable. Only full compounds are not sufficient for effective dictionary look-up and for searching compound words. Compound splitting with a morphological analyzer and separate translation of component words are useful for getting good performance in retrieval. In dictionary-based approach one of the factors that cause errors is Multi-term phrases translation [4]. In [5] proposed word-by-word translation using a dictionary that causes performance degradation than manual phrases translation. Identifying phrases in the query and translating them using a phrase look-up dictionary can reduce the impact of the phrases and compound words translation. It is unable to create "complete" phrase dictionary because using a language is a creative activity and new phrases are added continuously. If a phrase is not stored in a lexicon then identifying it in a query and translation remains as a major problem.

4 Translation Ambiguity

The effectiveness of CLIR will be 60% lower than monolingual retrieval if the problem of translation ambiguity is not considered in simple dictionary translation [6], [7]. Translation ambiguity is caused due to some words that may have different translations and meanings. This problem is observed when short queries are entered by the web users. In the dictionary-based approaches, most probably the errors are caused due to selecting the wrong translation terms among the translations provided by the dictionary. There are different techniques to reduce the ambiguity and errors caused during query translation. In machine readable dictionaries instead of using all possible translations, we can use the techniques based on term co-occurrence [6], [8], term similarity [9], [10] and language modelling [11]. Use a general and a domain-specific dictionary to combine the results of translating query terms and then use structural tags to indicate the contextual relationship among the resulting terms, which is a different approach to reduce the ambiguity problem that was proposed in [12]. We can also use pseudo-relevance feedback (PRF) to solve the ambiguity. In term co-occurrence, [6] a technique was described that employs co-occurrence statistics obtained from the corpus being searched to disambiguate dictionary translation. Their hypothesis is that the correct translation of query terms should co-occur in target language documents and incorrect translation should not co-occur. Term similarity is based on the concept of statistical term similarity. In [9] proposed a translation disambiguation technique. This technique selects the best Indonesian translation of an English term from all possible translations given by a bilingual dictionary. It makes use of term-similarity matrix in order to build the statistical term-

distribution parameters obtained for Indonesian terms taken from Indonesian corpus and a subset of their English collections for English terms. In Language modelling for dictionary-based query translation, a novel statistical model which uses hidden Markov model (HMM) [13] for selecting translations [11]. Probability of any query-translation pair will be computed by this query-translation model.

In [14] suggested that the most important step in Cross-Lingual Word Sense Disambiguation (CLWSD) is that a system has to choose the dictionary that provides the possible translations. In this a comparison is done between different dictionaries. It is based on two different frameworks. Firstly, by using theses dictionaries for an ideal system we analyze the potential results. In the other framework results are analyzed using different bilingual dictionaries. It also considers particular unsupervised CLWSD system CO-Graph. The analysis performed on the dictionaries for a particular system has improved the results in that framework. In [15] suggested that in order to assign the most appropriate sense to a polysemous word, Word sense disambiguation (WSD) is used: a raw text corpus and a machine-readable dictionary (MRD) are the two resources for automatic WSD, using which a method is proposed. For every occurrence of polysemous words in a sentence the system separately constructs the acyclic weighted digraph (AWD) that disambiguate all occurrences of polysemous words in a sentence. In [16] suggested that in Cross Language Information Retrieval one common way to translate the query is to use bilingual dictionaries. Dictionary-based query translation is improved by using Vietnamese-English Bilingual Information Retrieval and it may also consists of algorithms for query segmentation, word disambiguation and re-ranking. The proposed algorithms are compared and verified with the baseline method by implementing an evaluation environment.

In [17] proposed that in cross-language information retrieval, a method using dynamic incremental clustering is proposed to resolve ambiguities implicitly. In translated quires, document clusters can resolve the ambiguities efficiently and also considers the context of all the terms into account in a document. By using vector space retrieval model or the probabilistic retrieval model for translated query terms we can retrieve the documents. For this a framework has been proposed that translates a query in Korean/Japanese into English by looking up bilingual dictionaries. For the top-ranked retrieved documents, query-oriented document clusters are incrementally created and the weight of each retrieved document is recalculated by using the clusters. In [18] proposed that in CLIR for Dictionary-based query translation a source term can yield term having different meanings. Based on target document collection we check the methods that solve the ambiguity of translations. The two kinds of disambiguation technique are term co-occurrence statistics and pseudo-relevance feedback. The CLEF 2003 test collection for German to Italian bilingual searches are used to compare these techniques. The experiments showed that the term co-occurrence based techniques is dominant and the PRF method shows high search performance, but the statistical tests does not support these conclusions. In [19] suggested that in CLIR the crucial process is to solve the ambiguity in the translation of short length queries. If we use a bilingual dictionary this problem becomes more challengeable. In this for dictionary-based CLIR a statistical framework is developed by using monolingual word co-occurrence statistics. In this we can estimate translation probabilities of query words. When we compare the proposed work with the previous work on dictionary-based CLIR, the advantages identified are to capture the uncertainty in translating queries by calculating the translation probabilities explicitly, simultaneous estimation for the translations of all query words and with a unique optimal solution a formulated problem can be solved.

In [20] suggested that in cross-language information retrieval, query translation is an effective method. In this multiple translations of query terms causes the problem of translation ambiguities. In this we propose a method for source query terms that examine all combinations of target query term. This method causes some problems while using bilingual dictionary for CLIR with query translation. So the associations between the translation equivalent at hand and all other translation equivalents of adjacent query terms are considered instead of selecting the best translation equivalent of a query term. In [21] proposed that using binary relevance assessments, cross-language information retrieval has been restricted to settings. In a best match retrieval environment using graded relevance assessments, dictionary-based CLIR results are evaluated. For testing, a text database containing newspaper articles and its related topics are considered. From the topics, monolingual baseline queries will be formed automatically; next translation of source language topics into target language will be done automatically. A comparison to measure the effectiveness of translated queries to that of monolingual queries was performed. Later, to expand the original target queries, pseudo-relevance feedback was used. Using stringent, regular and liberal relevance thresholds, CLIR performance was evaluated. The regular or liberal threshold yielded reasonable performance, where stringent threshold could not achieved high performance. Pseudo-relevance feedback based query expansion successfully improved the performance of the translated queries on all the relevance thresholds. The stringent threshold performance has not improved by this method in relation to the other thresholds. In [22] proposed that for many languages, lack of comprehensive dictionaries becomes a bottleneck in dictionary-based CLIR. A method is proposed using simple seed lexicons where the multilingual dictionaries (for Spanish and Swedish) will automatically emerge. By cognate mapping the seed lexicons will be emerged automatically. Validate Lexical and semantic hypotheses and then iteratively generate new ones by making use of co-occurrence patterns in parallel corpora, which are hypothesized translation synonyms. By using a large medical document collection, evaluate these newly derived dictionaries within a cross-language retrieval setting.

5 Out-Of-Vocabulary Terms

The Out-Of-Vocabulary (OOV) problem is caused due to the missing of query terms in bilingual dictionaries and parallel corpora. For example, if the query is regarding the current affairs it may contain new words that are not present in the

translation dictionary. This will degrade the performance of CLIR. A wide variety of solutions to the problem of OOV terms are present in existing CLIR systems. One of them is to use domain-specific bilingual dictionaries but they are costly to produce. This makes a move towards transliteration. In Transliteration, an important problem in CLIR is missing translations. This is caused when there is an unknown source term (the OOV problem) or the source term is known but the translation is missing. The OOV problems often concern proper names. When such a term is present in a query which gives vital information is not translated properly, it may result in low retrieval effectiveness. OOV term translation should be based on the sound rather than on the meaning otherwise it fails. In the process of transliteration, the source language original term is converted into a target language approximate phonetic equivalent.

In [23] suggested that to identify the translation equivalents of source words that are obtained by transformation rule based translation (TRT) a method based on the statistical technique was developed. Frequency-based identification of translation equivalents (FITE) effectiveness was tested using biological and medical cross-lingual spelling variants and OOV words in Spanish-English and Finnish-English TRT. The technique also reliably identified native source language words i.e. the source words that cannot be correctly translated by TRT. Rather than basic dictionary-based CLIR, the Dictionary-based CLIR augmented with FITE-TRT performs well, where OOV keys were kept intact. Among several fuzzy translation/matching approaches in CLIR experiments, the FITE-TRT with Web document frequencies was the best technique. In [24] suggested that in cross-language information retrieval, noise can present in queries or in the target collection. To translate queries, OOV for dictionaries is used. Similar problems are identified with historic document retrieval (HDR), OCR errors and historical spelling variants. To solve these problems three data driven approaches have been proposed. The first two methods that operate on string level are the transformation rule based translation (TRT) method and the classified s-gram method. In the target query, the query word which is to be included, is identified from the target document that is approximate match of a query word. In the third method, translation of OOV words requires translation knowledge. For this translation the corpus-based approach, parallel or comparable corpora are used. In [25] suggests the major problem in dictionary-based cross-language information retrieval is technical terms and proper names. In this for cross-lingual spelling variants a novel two-step fuzzy translation technique was proposed. Firstly, for source words we apply transformation rules to make them similar to their target language equivalents. By using translation dictionaries as source data we can generate rules automatically. In the next step, by using fuzzy matching the intermediate forms obtained in the first step are translated into a target language. In some cases the two-step technique outperformed than fuzzy matching alone.

In [26] suggested that in Natural Language Processing (NLP) one of the major tasks is Named Entity (NE) extraction. Complexity of aligning NEs in bilingual documents is more than that of identifying NEs in monolingual documents. With multiple knowledge sources, by incorporating statistical models a new model was developed for aligning bilingual NEs. In this proposed approach we translate an English NE phrase into a Chinese equivalent. For this, a word translation uses lexical translation and word reordering uses alignment probabilities. The method contains automatically learning phrase alignment, acquiring word translations and automatically discovering transliteration transformations. The method also contains language-specific knowledge functions. At run time, in a pair of bilingual sentences for each source NE we apply the proposed models to generate and evaluate the target NE candidates. Then the computed probabilities are used to align source and target NEs. In [27] suggested a Korean language processing Compound noun segmentation as a first step. In general human supervision is required for most the approaches that causes unknown word problem and these can be overcome by unsupervised approaches. Generally unsupervised methods depend on character-based segmentation clues and will not consider all possible segmentation candidates. To overcome the problem, we use an unsupervised segmentation algorithm that makes use of word-based segmentation context. In word-based segmentation clues, by using corpus we can generate dictionary automatically. Improvement in dictionary-based longest-matching algorithm proves that proposed experiments are successfully applied to Korean information retrieval.

Some others independent methods exist in the literature, other than discussed above which will improve the performance of dictionary based translation in CLIR. In [28] suggested that impressive results have been achieved by automatic construction of bilingual dictionaries. In general, parallel corpora is used in Bilingual dictionaries that makes use of selected text domains and language pairs. Due to this other potential resources are explored. For bilingual terminology extraction one can make use of Wikipedia as a corpus. From different types of Wikipedia links information, to extract term-translation a method has been proposed. To determine the correctness of unseen term-translation pairs an SVM classifier is trained on the features of manually labeled training data. In [29] proposed a multilingual parallel electronic dictionary, named MPEDM. It covers the languages in Chinese and Mongolian. By using Cosine similarity measure on different Mongolian systems the average coverage of MPEDM was evaluated. Later manually add a part-of-speech (POS) and Chinese word pairs to MPEDM and then the traditional and TODO Mongolian words were paralleled automatically. The MPEDM dictionary can be used in multilingual word searching and interpreting a word in its reading and grammatical form on line. In [30] proposed that patent retrieval is one of the branches of Information Retrieval (IR) which aims to support patent professionals in retrieving patents. In order to reduce the difficulty of accessing the patents by their language, they have to be translated into other languages. This improves the opportunities of patent retrieval and it is exploited by query translation. By using general domain-free dictionary and a domain-specific patent dictionary we can expand the translations of monolingual patent quires. In [31] suggested that in natural language (NL) dictionaries a Trie structure is a frequently used approach for retrieval. Computer hard disk has a huge amount of common information in some dictionaries

when we develop a variety of NL processing systems. In this a method is proposed to merge both individual dictionaries and generalized dictionary. It decreases the size of total dictionary and increases the usage of individual dictionaries to that of the other applications.

In [32] suggested that by using new translations in comparable corpora the Bilingual dictionaries are extended automatically. "Similar words have similar contexts" is the basic assumption. In this we extract the salient pivot words for which translation is available in bilingual dictionary by using Bayesian estimation, which addresses the issues based on the context of a word with an unknown translation (query word). Later, for a query word we match the pivot words to identify their translations. Now by using probability, a similarity score is calculated between the translation candidate and query word. In [33] suggested that a transitive translation is a branch of cross-language information retrieval in which search queries are translated into the document language by using intermediate (or pivot) language. In the experiments, queries constructed from CLEF 2000 and 2001 Swedish, Finnish and German topics were translated into English through Finnish and Swedish by an automated translation process using morphological analyzers, stop-word lists, electronic dictionaries, n-gramming of untranslatable words, structured and unstructured queries. The results of the transitive runs and bilingual runs were compared. The transitive runs using structured target queries performed well.

In [34] suggested that we can identify the important issues in dictionary based CLIR .To explain the relationships between existing techniques, unified frameworks for term selection and term translation are explored. Using unifom query translation architecture we can illustrate the effect of those techniques with the help of four contrasting languages for systematic experiments. In the key results we have identified the previously unseen dependence of pre- and post-translation expansions. In [35] proposed that by depending on the user information needs Query suggestion helps to suggest relevant queries for a given query. In this cross-lingual query suggestion (CLQS) for a query is used to suggest relevant queries in other languages. In the query log CLQS presents an effective means of mapping the input query of one language to queries of the other language. With a discriminative model we can estimate the cross-lingual query similarity. Benchmarks have shown that the CLQS system outperforms a baseline system that makes use of dictionary-based query translation. In [36] suggested that in large scanned book collections we can identify translations of books with OCR errors. In this method linear progression of ideas in a book is preserved. Taking into account English and German language books an English-German dictionary is proposed to translate the word sequence of English book into German. By using Longest Common Subsequence (LCS) algorithm both sequences can be aligned as they are in German. In [37] proposed that in order to identify predefined entities from a document, Dictionary-based entity extraction is used. Approximate entity extraction is used to improve extraction recall. It will find all the substrings in the document that matches entities approximately in a given dictionary. To support many similarity/dissimilarity functions a unified framework is proposed. Some of the similarity/dissimilarity functions are jaccard similarity, cosine similarity, dice similarity, edit similarity and edit distance. To utilize the shared computation, efficient filtering algorithms are developed and effective pruning techniques are developed. When this method is compared with state-of-the-art studies it has greater performance.

6 Conclusion

CLIR is used for retrieving the information in Languages other than native languages. When user needs the information in other language we can make use of CLIR system. Dictionary based CLIR is mostly used in information retrieval and machine translation. In dictionary based CLIR we perform query translation by using bilingual dictionaries. This causes the problems like selection of translation for query, selection of the dictionary for possible translation and so-on. To solve dictionary selection problems we can use dictionaries like multilingual parallel electronic dictionary and by using compression algorithm we can reduce the size of dictionary. Performance enhancement of the dictionary based CLIR mainly depends on three factors such as phrases and compound words, translation ambiguity, and out-of-vocabulary terms. Translation ambiguity problems can be solved by using techniques like Word Sense Disambiguation, heap-based filtering algorithms, document clusters, examining all combinations of target query term translations corresponding to the source query terms, term co-occurrence based method, PRF based method, graded relevance assessments and statistical framework. Out-of-vocabulary problems can be solved by using the techniques like novel two-step fuzzy translation technique, translation identification framework, novel statistical frequency-based identification of translation equivalents, unsupervised segmentation algorithm, using the three methods: transformation rule based translation method, the classified s-gram method and the corpus-based approach. Transliteration problems can be solved by using techniques like lexical translation/transliteration. For improving the performance of retrieval by using pivot language we can use techniques like Bayesian estimation, automated translation process, combining two bilingual dictionaries. By mining relevant queries in different languages from query logs also we can improve the performance of dictionary based CLIR. So we have seen insight into different techniques to solve the problems and improve the performance of dictionary based CLIR. According to research studies made till now on Dictionary based translation, WSD systems achieve sufficiently high levels of accuracy on a variety of word types and ambiguities.

References

[1] Jian-Yun Nie, Cross-Language Information Retrieval. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010.

[2] Francis Bond, and Kentaro Ogura, "Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary," Language Resources and Evaluation, Volume 42, Issue 2, 127-136, 2007.

[3] Eija Airio, and Kimmo Kettunen, "Does dictionary based bilingual retrieval work in a non-normalized index?," Information Processing and Management,

45, 703–713, 2009.

[4] Lisa Ballesteros, and W. Bruce Croft, "Phrasal translation and query expansion techniques for cross-language information retrieval," In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 84-91, 1997.

[5] David A. Hull, and G. Grefenstette, "Querying across languages: a dictionary-based approach to multilingual information retrieval," In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 49-57, 1996.

[6] Lisa Ballesteros, &W. Bruce Croft. (1998). Resolving ambiguity for cross-language retrieval. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 64-71.

[7] Mohammed Aljlayl, and Ophir Frieder, "Effective Arabic-English cross-language information retrieval via Machine-readable dictionaries and machine translation," In Proceedings of the 10th International Conference on Information and Knowledge Management, 295-302, 2001.

[8] Jianfeng Gao, Ming Zhou, Jian-Yun Nie, Hongzhao He, and Weijun Chen, "Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations," In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 183-190, 2002.

[9] Mirna Adriani, "Using statistical term similarity for sense disambiguation in cross-language information retrieval," Information Retrieval, Volume 2, Issue 1, 71-82, 2000.

[10] Akira Maeda, Fatiha Sadat, Masatoshi Yoshikawa, and Shunsuke Uemura, "Query term disambiguation for web cross-language information retrieval using a search engine," In Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages, pages 25-32, 2000.

[11] Marcello Federico, and Nicola Bertoldi, "Statistical cross-language information retrieval using n-best query translations," In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 167-174, 2002.

[12] Ari Pirkola, "The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval," In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 55-63, 1998.

[13] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Readings in speech recognition, 267-296, 1990.

[14] Andres Duque, Juan Martinez-Romo, and Lourdes Araujo, Choosing the best dictionary for Cross-Lingual Word Sense Disambiguation," Knowledge-Based Systems, 81, 65–75, 2015.

[15] Yeohoon Yoon, Choong-Nyoung Seon, Songwook Lee, and Jungyun Seo, "Unsupervised word sense disambiguation for Korean through the acyclic weighted digraph using corpus and dictionary," Information Processing and Management, 42, 710–722, 2006.

[16] Lam Tung Giang, Vo Trung Hung, and Huynh Cong Phap, "Experiments with Query Translation and Re-ranking Methods in Vietnamese-English Bilingual Information Retrieval," SoICT '13 Proceedings of the Fourth Symposium on Information and Communication Technology, 118-122, 2013.

[17] Kyung-Soon Lee, Kyo Kageura, and Key-Sun Choi, Implicit ambiguity resolution using incremental clustering in cross-language information retrieval," Information Processing and Management: an International Journal, Volume 40, Issue 1, 145–159, 2004.

[18] Kazuaki Kishida, "Term disambiguation techniques based on target document collection for cross-language information retrieval: An empirical comparison of performance between techniques," Information Processing and Management, 43, 103–120, 2007.

[19] Yi Liu, Rong Jin, and Joyce Y. Chai, "A Statistical Framework for Query Translation Disambiguation," ACM Transactions on Asian Language Information Processing, Vol. 5, No. 4, 360–387, 2006.

[20] Hee-Cheol Seo, Sang-Bum Kim, Hae-Chang Rim, and Sung-Hyon Myaeng, "Improving query translation in English–Korean cross-language information retrieval". Information Processing and Management, 41, 507–522, 2005.

[21] Raija Lehtokangas, Heikki Keskustalo, and Kalervo Järvelin, "Experiments with dictionary-based CLIR using graded relevance assessments: Improving effectiveness by pseudo-relevance feedback," Information Retrieval, Volume 9, Issue 4, 421-433, 2006.

[22] Korne'l Marko´, Stefan Schulz, Olena Medelyan, and Udo Hahn, "Bootstrapping Dictionaries for Cross-Language Information Retrieval," SIGIR '05 Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 528-535, 2005.

[23] Ari Pirkola, Jarmo Toivonen, Heikki Keskustalo, and Kalervo Järvelin, "Frequency-Based Identification of Correct Translation Equivalents (FITE) Obtained Through Transformation Rules," ACM Transactions on Information Systems, Vol. 26, No. 1, Article 2, 2007.

[24] Antti Järvelin, Tuomas Talvensaari, and Anni Järvelin, "Data Driven Methods for improving Mono- and Cross-lingual IR Performance in Noisy Environments," AND '08 Proceedings of the second workshop on Analytics for noisy unstructured text data, 75-82, 2008

[25] Jarmo Toivonen, Ari Pirkola, Heikki Keskustalo, Kari Visala, and Kalervo Järvelin, "Translating cross-lingual spelling variants using transformation rules," Information Processing and Management, 41, 859–872, 2005.

[26] Chun-Jen Lee, Jason S. Chang, and Jyh-Shing R. Jang, "Alignment of Bilingual Named Entities in Parallel Corpora Using Statistical Models and Multiple Knowledge Sources," ACM Transactions on Asian Language Information Processing, Vol. 5, No. 2, 121-145, 2006.

[27] In-Su Kang, Seung-Hoon Na, and Jong-Hyeok Lee, "Collection-based compound noun segmentation for Korean information retrieval," Information Retrieval, Volume 9, Issue 5, 613-631, 1 September 2006.

[28] Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio, "Improving the Extraction of Bilingual Terminology from Wikipedia," ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 5, No. 4, Article 31, 2009.

[29] I. Dawa, W. Aishan, and B.Dorjiceren, "Design and Analysis of a POS Tag Multilingual Dictionary for Mongolian," IERI Procedia, 7, 102 – 112, 2014.

[30] Charles Jochim, Christina Lioma, Hinrich Schütze, Steffen Koch, and Thomas Ertl, "Preliminary Study into Query Translation for Patent Retrieval," In Proceeding PaIR '10 Proceedings of the 3rd international workshop on Patent information retrieval, 57-66, 2010.

[31] Y. Kadoya, M. Fuketa, El-Sayed Atlam, K. Morita, T. Sumitomo, and J. Aoe, "A compression algorithm using integrated record information for translation dictionaries," Information Sciences—Informatics and Computer Science: An International Journal - Special issue: Informatics and computer science intelligent systems applications, Volume 165, Issue 3-4, 171–186, 2004.

[32] Daniel Andrade Takuya Matsuzaki, and Jun'ichi Tsujii, "Statistical Extraction and Comparison of Pivot Words for Bilingual Lexicon Extension," ACM Transactions on Asian Language Information Processing, Vol. 11, No. 2, Article 6, 2012.

[33] Raija Lehtokangas, Eija Airio, and Kalervo Järvelin, "Transitive dictionary translation challenges direct dictionary translation in CLIR," Information Processing and Management: an International Journal, Volume 40, Issue 6, 973–988, 2004.

[34] Gina-Anne Levow, Douglas W. Oard, and Philip Resnik, "Dictionary-based techniques for cross-language information retrieval," Information Processing and Management, 41, 523–547, 2005.

[35] Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Kam-Fai Wong, and Hsiao-Wuen Hon, "Exploiting Query Logs for Cross-Lingual Query Suggestion,"

ACM Transactions on Information Systems, Vol. 28, No. 2, Article 6, 2010.

[36] Ismet Zeki Yalniz, and R. Manmatha, "Finding Translations in Scanned Book Collections," SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 465-474, 2012.

[37] Guoliang Li, Dong Deng, and Jianhua Feng, "Faerie: Efficient Filtering Algorithms for Approximate Dictionary-based Entity Extraction," SIGMOD '11 Proceedings of the ACM SIGMOD International Conference on Management of data, 529-540, 2011.