Cross-lingual Information Retrieval

State-of-the-Art

Nurul Amelina Nasharuddin, Muhamad Taufik Abdullah
Department of Multimedia
Faculty of Computer Science & Information Technology, UPM
43400 UPM Serdang, Serdang, Selangor, Malaysia
nurulamelina@gmail.com, taufik@fsktm.upm.edu.my

***Abstract* – Information retrieval involves finding some required information in a collection of information or in databases. The information or database need not necessarily be in one language. In other words, language should not limit the finding of information. The way to search for the information is by looking at every item in the collection and when the need to translate the language arises, the techniques and methods developed for the cross-lingual retrieval system is used. This paper reviews some recent researches focusing on topics in cross-lingual information retrieval and their role in current research directions which include new models and paradigms in the wide area of information retrieval.**

Keywords – cross-lingual information retrieval; query translation; document translation

I. INTRODUCTION

The area of information access has evolved to include many sophisticated tasks such as information retrieval, question answering tasks, summarization, multimedia information retrieval, text mining, text clustering and web information retrieval. Information retrieval (IR) is “the act of finding materials, usually documents of an unstructured form that satisfies an information need within large collections stored in computers”[1]. These tasks are not restricted to only documents in one language but also in multiple languages. The classical IR normally regards the documents in foreign language as unwanted “noise” [2]. The need for handling multiple languages, introduce a new area of IR which takes into account all documents regardless of the languages being used taking into account cross-lingual and multi-lingual aspects. Whilst in classical IR search engines, both query language and the retrieved documents language are the same, in cross-lingual IR system, they can be different. In enhanced version of cross-lingual IR, where the retrieved documents are of multiple languages, there are many problems that can arise in implementing it. This paper will focus on the challenges and current approaches to overcome these problems.

Cross-lingual IR has become more important in recent years. Currently, searching which is a classical IR has been the most used tool in the Web but there are few satisfactory quality cross-lingual IR systems available for the web [3]. However, cross-lingual approaches for restricted domain such as in medicine and geo-informatics have shown to be popular. The basic idea behind the cross-lingual IR is to retrieve documents in a target language different from the query or source language. This may be possible even when the user querying is not a speaker of the language in the retrieved documents. The retrieved documents in different languages can later be translated by a human translator for use of the user. For example, when the user of a cross-lingual IR searches for the information about “traditional dress” the information about “kebaya” in the Malay language is also retrieved. The document in the Malays language can then be translated manually into English. Translations can actually be performed on the query, the document or in both document and query [4]. Query translation involves translating the query to the target language while document translation deals with translating the document into the source language.

At present, a number of tracks and workshops have been introduced to support research in cross-lingual IR. Cross-language Evaluation Forum (CLEF) is a forum that promotes research in multilingual system development since 2000 and deals mainly with European languages. The NTCIR (NII Test Collection for IR System) workshop is designed to enhance researches in cross-lingual IR, mainly in Japanese and other Asian languages. Cross-language retrieval track was offered at Text Retrieval Conference (TREC) only up to the year 2000 but it is still being studied in both CLEF and NTCIR. This paper will discuss further on the approaches of query and document translation, challenges in cross-lingual IR and current approaches to tackle these challenges. It is organized as follows; section II discusses query translation. Document translation is discussed in section III followed by the challenges in section IV. Current approaches is next in section V followed by the conclusion section.

II. QUERY TRANSLATION

Query translation can be based on using bilingual dictionary or using the corpora or machine translation. In dictionary-based query translation, the query will be processed linguistically and only the keywords are translated using Machine Readable Dictionaries (MRD). MRDs are electronic versions of printed dictionaries, either in the general domain or specific domain. The use of existing linguistics resources, especially the MRDs, is a natural approach to cross-lingual IR. Translating the query using the dictionaries is much faster and simpler than translating the documents [5]. The drawbacks of query translation are due to ambiguities, problems of word inflection and problems of translating word compounds, phrases, spelling variants and special terms. These will be discussed further in Section IV.

In Query translation using corpora, a corpus or a number of corpuses is used. Corpora, (plural of corpus) are the repositories of a collection of natural language material, such as texts, paragraphs and sentences from one or many languages. There are two types of bilingual corpora; parallel and comparable corpora [6]. Parallel corpora contain the same documents in more than one language. Aligned parallel corpus is annotated to show exactly which sentence of the source language corresponds exactly with the sentence of the target text. They can be used in analyzing many processes involved in transferring information, ideas, and concepts from one language to another. They also become the sources of translation equivalent data for human or machine translation applications. Comparable corpora are sets of documents in multiple languages which can be compared not because they are translations of each others, but because they cover the same area and therefore contain an equivalent vocabulary [1]. A good example is the multilingual news feeds produced by news agencies such as Reuters, CNN, BBC, Xinhua News and BERNAMA. Such texts are widely available on the Web for many language pairs and domains. They often contain many sentence pairs that are fairly good translations of each other [7].

Cross-lingual IR with query translation using machine translation seems to be an obvious choice [8] compared to the other two above. Machine translation has proved to be an elusive goal, but today a number of systems are available which produce output which, though not perfect, is of sufficient quality to be useful in a number of specific domains. The advantages of using the machine translation is that it saves time while translating large texts. Cross-lingual IR is difficult if the translation is based on machine translation alone. Queries that the user enters to the search engine are often short thus provide little context for word disambiguation. Another factor that contribute to the difficulties of machine translation is in handling the grammar [9]. To overcome these, some systems use the dictionary-based methods or the corpora.

One important approach in addition to query translation is by adding translated queries with relevant terms. This has been shown to improve cross-lingual IR effectiveness [10]. The approach is known as query expansion and can be done by expanding the query both before and after translation. Expanding the query before the translation results in including more terms in the query language and doing the expansion after translation reduces the effect of irrelevant query terms by adding more context specific terms. One of the query expansion techniques is called the pseudo relevance feedback [11, 12]. This technique is based on an assumption that the top few documents initially retrieved are indeed relevant to the query, and so they must contain other terms that are also relevant to the query. The query expansion technique adds such terms into the previous query. The tf*idf term weighting formula will choose the relevant terms from the top ranked documents and a certain number of terms that have the highest weight scores will be added to the previous query [1].

III. DOCUMENT TRANSLATION

Normally document translation is typically done using a machine translation system, such as the SYSTRAN [13], PROMPT[16] and AppTek [14]. Machine translation is the standard name for computerized systems that produce translations of natural languages, with or without human assistance. The basic tasks of machine translation system can be generalized as source text analysis, source-target transfer, and target language generation in conjunction with bi or multi-lingual dictionaries [15]. Morphological, syntactic, and semantic information are accumulated and recorded throughout the entire process.

SYSTRAN is one of the oldest machine translation companies. US government funded SYSTRAN in 1995 for developing a cross-language information retrieval system based on its natural language parsing and machine translation technology. SYSTRAN's software which combines the rule-based and statistical machine translation delivers high quality translation for any domain [13]. SYSTRAN's engine reduces the amount of data required to train the software and the size of the statistical models. The statistical technique learns from existing monolingual and bilingual collection to improve the translation process. PROMPT is another provider of machine translation technology. It provides some translation solutions such as the machine translation systems and services, dictionaries, translation memory systems, and also data mining systems [16]. PROMPT machine translation provides solutions for issues of dictionary volume and translation modules, and offers additional software tools for creating and editing for dictionaries, linguistic editor, interface, and post-editing tools.

In 1999, McCarley pointed several possible advantages of document translation and there are more chances for translating a word correctly. Researches in comparing both translation approaches showed that document translation is typically better than query translation [17, 18]. Apart from what have being discussed in Section II, there are some drawbacks of using the document translation. Machine translation is computationally expensive and sometimes impractical. However, with modern computers, this is becoming less of a problem, especially for smaller document collections. Among the other problems are the cost of machine translation systems and the unavailability of translation systems for a wide range of language pairs. Obviously, translating the query terms is more practical, since entire collections of documents may be very large and out-of-date to be translated [19]. However, with the query-translation approach, ambiguity can be a serious problem.

IV. CHALLENGES IN CROSS-LANGUAGE IR

Each of the approaches in listed in Section II and III has created challenges to the cross-lingual IR system. One of the problems is the translation disambiguation, which often rooted from homonymy and polysemy [2]. Homonymy refers to a word that has at least two entirely different meanings, for example the word “left” can either mean the opposite of right or the past tense of leave. Polysemy refers to a word which can take on two distinct but related meanings such as the “head” of the family or the human’s “head”. It becomes a problem when finding the most appropriate translation from several choices in the dictionary. Very frequently, translation of a word results in such a choice having to be made. For example, the Malay word “rajah” has many different translations into English, such as “chart”, “diagram”, “tattoo” and “amulet” to name a few [20].

A common problem with query translation is word inflection used in the query. This can be solved by stemming and lemmatization [1]. Lemmatization is where every word is simplified to its uninflected form or lemma; while stemming is where different grammatical forms of a word are reduced to a common shorter form called a stem, by removing the word endings. For example, the stemming rules for word “see” might return just “s” by stemming and “see” or “saw” by lemmatization.

Using the dictionary-based translation is a traditional approach in cross-lingual IR systems but significant performance degradation is observed when queries contain words or phrases that do not appear in the dictionary. This is called the Out-of-Vocabulary (OOV) problems [21]. This is to be expected even in the best of dictionaries. Input queries by user usually short and even the query expansion cannot help to recover the missing words because of the lacking information. Generally, OOV terms are proper names or newly created words. For example, a user wants to search the information about the Influenza A (H1N1) disease in Malaysia by entering “H1N1 Malaysia” as the query. The H1N1 is a newly created term and may not be included in a dictionary which was published only a few years ago. If the term H1N1 is omitted from the query translation, it is most likely that the user will not get any relevant documents at all. OOV terms include compound words, proper nouns and technical terms [22].

In many documents, technical terms and proper names are important text elements. Dictionaries only include the most commonly used proper nouns and technical terms used such as major cities and countries. Their translation is crucial for a good cross-language IR system. A common method used to handle untranslatable keywords is to include the untranslated word in the target language query. If this word does not exist in the target language, the query will be less likely to retrieve the relevant documents. Translating phrases is also becoming one of the problems in cross-lingual IR. A phrase cannot be translated by translating each of the word in the phrases [21]. An example is the idiom which when translating word by word, the meaning will be totally different from the actual meaning in source language.

Named entities (NEs) are essential components of texts, especially news texts [23]. NEs extraction and translation are vital in the field of natural language processing (NLP) for research on machine translation, cross-language IR, bilingual lexicon construction, and so on. There are three types of NEs [24]; entity names (organizations, persons and locations), temporal expressions (dates and times), and number expressions (monetary values and percentages). Organizations, person and location named entities are difficult to handle with a fixed set of rules, since new entity names are constantly being created, and hence the growing need to investigate techniques for NEs extraction and translation. Bilingual dictionaries often have few entries for NEs [19]. But, when NEs are wrongly segmented as ordinary words and translated with a bilingual dictionary, the results are often poor.

V. CURRENT APPROACHES

Wikipedia [25] has become an important resource in the cross-lingual IR recently. Many researchers have conducted studies and experiments using the free online encyclopedia. Lin et al [19] have developed a Japanese-Chinese IR system based on the query translation approach. The system employed a more conventional Japanese-Chinese bilingual dictionary and Wikipedia for translating query terms. They studied the effects of using Wikipedia and proposed that Wikipedia can be used as a good NEs bilingual dictionary. By exploiting the nature of Japanese writing system, the query terms were processed differently based on the forms they were written in. To cope with term disambiguation, they have adopted an iterative disambiguating method based on the PageRank algorithm. The method proved to be effective and outperformed the previous Japanese-Chinese systems’ tests.

A recent Wikipedia-based study by Nguyen, Overwijk, Hauff, Trieschnigg, Hiemstra and de Jong [26] showed that query translations for cross-lingual IR can be performed using only Wikipedia. An advantage of using Wikipedia is that it allows translating phrases and proper nouns well. It is also very scalable since it is easy to use the most up-to-date version of Wikipedia which makes it able to handle actual terms. The approach is that the queries are mapped to Wikipedia concepts and the corresponding translations of these concepts in the target language are used to create the final query. WikiTranslate system [26] is evaluated by searching the topics in Dutch, French, and Spanish language within an English data collection. The system which achieved a performance of 67% compared to the monolingual baseline can be a valuable alternative to current translation resources. The unique structure of Wikipedia (for example the text and internal links) can be very useful in cross-lingual IR. The use of Wikipedia might also be suitable for interactive cross-lingual IR, where user feedbacks are also used to translate the query, since Wikipedia is already very popular among internet users.

Query suggestions aim to suggest relevant queries for a given query, which help users to specify their information needs better [27]. It is closely related to query expansion but query suggestions will suggest full queries that have been formulated by users in another language. Gao et al [27] proposed query suggestions by mining relevant queries in different languages from up-to-date query logs as it is expected that for most user queries, we can find common formulations on these topics in the query log in the target language. Therefore, cross-lingual query suggestions also play a role of adapting the original query formulation to the common formulations of similar topics in the target language. Used as a query translation system, the proposed method demonstrates higher effectiveness than traditional query translation methods using either bilingual dictionary or machine translation tools.

Pourmahmod and Shamsfard in [28] carried out a research to retrieve English documents relevant to Persian queries using bilingual ontologies to annotate the documents and queries. A bilingual ontology consists of ontology and a bilingual dictionary. Ontology is a formal, explicit specification of a shared conceptualization. It contains a set of distinct and identified concepts related by a set of relations [29]. They used the ontology to expand the query with related terms in pre- and post-translation expansion and the combined approach significantly improves cross-lingual performance.

Researchers in [3] analyzed the query translation in cross-lingual IR based on feature vectors and usage of context information during the query translation. They pointed out that by using information external to the query, such as the ontologies and document collections, the effects of disambiguation and polysemy can be reduced. The characteristics of a feature vector are dependent on the quality of both the ontology and the document collection being used. As the research is still in progress, they still need to fully implement the approach for more thorough testing and evaluation. But an advantage of this approach is the adaptability to several languages, which can be done by adding other dictionaries and thesauruses.

Disambiguation is the aim of most translation techniques used in cross-language IR. Yuan and Yu [30] found a method using co-occurrences between pairs of terms as statistical measure, unlike the traditional statistical approach. This method needs only a bilingual dictionary and a monolingual corpus for translation. They compared different combinations of target terms and presented the output in the form of probability distribution. Using the results, the query is converted to target language. It is a simple method and experiment showed that it performed well.

The increasing numbers of multi-lingual documents in Web posed a challenge in managing them. Wu and Lu [31] identifies a novel model called domain alignment translation model to conduct cross-lingual document clustering and term translation simultaneously and in the end the multi-lingual documents with similar topics can be clustered together. Their method with the use of only a bilingual dictionary can achieve comparable performance with the machine translation method using Google translation tool. Although their experiments only consider word but ignoring the base phrase, the clustering in the source language and the clustering in the target language are related highly and the clustering quality can be emphasized for future research.

VI. CONCLUSION

Cross-lingual IR provides new paradigms for searching documents through myriad varieties of languages across the world and it can be the baseline for searching not only among two languages but also in multiple languages. This paper explains a description on cross-lingual IR, its challenges and current methods and techniques to overcome problems for efficient and resourceful searching. The purpose of this paper is to review some of the latest researches in the area of cross-lingual IR. Survey indicates that query translation is always the choice as compared to document translation. It is more convenient to translate only the query than the whole documents. Document translation which uses machine translation is computationally expensive and the size of document collection is large. However, it might be practical in the future when the computer technology improves.

In Malaysia, there are currently researches conducted on cross-lingual IR especially on the Malay-English cross-lingual IR. However, the Malay language retrieval systems are still lagging behind as compared to researches on other languages such as English and European languages. It is hoped that more researches that focus on the Malaysian languages will be conducted in the future.

REFERENCES

[1] D. Manning, C., P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*. 2009, Cambridge: Cambridge University Press.

[2] Abusalah, M., J. Tait, and M. Oakes (2005) *Literature Review of Cross Language Information Retrieval*. World Academy of Science, Engineering and Technology 4, 175-177.

[3] Lilleng, J. and S.L. Tomassen, *Cross-lingual information retrieval by feature vectors*, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2007. p. 229-239.

[4] Ren, F. and D.B. Bracewell, *Advanced Information Retrieval*. Electronic Notes in Theoretical Computer Science, 2009. 225: p. 303-317.

[5] McCarley, J.S., *Should we translate the documents or the queries in cross-language information retrieval?*, in *Proceedings of the 37th annual meeting of the Association*

for Computational Linguistics on Computational Linguistics. 1999, Association for Computational Linguistics: College Park, Maryland.

[6] Picchi, E. and C. Peters, *Cross-language information retrieval: a system for comparable corpus querying*, in *Cross-Language Information Retrieval*, G. Grefenstette, Editor. 2000, Kluwer Academic Publishing: Massachusetts. p. 81-90.

[7] Munteanu, D.S. and D. Marcu. *Extracting parallel sub-sentential fragments from non-parallel corpora*. in *Proceedings of the 21st international Conference on Computational Linguistics*. 2005. Sydney, Australia: Association of Computational Linguistics.

[8] Zhang, T. and Y. Zhang. *Research on chinese-english information retrieval*. in *Proceedings of the Seventh International Conference on Machine Learning and Cybernetics*. 2008. Kunming, China.

[9] Aljlayl, M. and O. Frieder, *Effective arabic-english cross-language information retrieval via machine-readable dictionaries and machine translation*, in *Proceedings of the tenth international conference on Information and knowledge management*. 2001, ACM: Atlanta, Georgia, USA.

[10] Xu, J. and W.B. Croft, *Query expansion using local and global document analysis*, in *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. 1996, ACM: Zurich, Switzerland.

[11] Attar, R. and A.S. Fraenkel, *Local feedback in full-text retrieval systems.* Journal of the Association for Computing Machinery, 1977. 24: p. 397-417.

[12] Ballesteros, L. and W.B. Croft. *Resolving ambiguity for cross-language retrieval*. in *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1998. New York: ACM Press.

[13] Anonymous (2002) *The Systran Linguistics Platform: A Software Solution to Manage Multilingual Corporate Knowledge*.

[14] Boretz, A., *AppTek Launches Hybrid Machine Translation Software*, in *Speech Tag Online Magazine*. 2009.

[15] Gachot, D.A., E. Lange, and J. Yang, *The SYSTRAN NLP browser: an application of machine translation technology in cross-language information retrieval*, in *Cross-Language Information Retrieval*, G. Grefenstette, Editor. 2000, Kluwer Academic Publishing: Massachusetts. p. 105-118.

[16] Anonymous (2005) *A Brief Guide to PROMPT Machine Translation Technology*.

[17] 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*. 1998, Springer-Verlag. p. 472-483.

[18] Chen, A. and F.C. Gey, *Combining Query Translation and Document Translation in Cross-Language Retrieval*, in *Comparative Evaluation of Multilingual Information Access Systems*. 2004, Springer Berlin: Heidelberg. p. 108-121.

[19] Lin, C.-C., et al., *Learning Weights for Translation Candidates in Japanese–Chinese Information Retrieval.* Expert Systems with Applications, 2009. 36(4): p. 7695-7699.

[20] Anonymous, *Rajah*, in *Kamus Daya*, C.S. Huat and L.Y. Choy, Editors. 2008, Penerbitan Minda (M) Sdn. Bhd.: Seri Kembangan.

[21] Lu, C., Y. Xu, and S. Geva, *Translation disambiguation in web-based translation extraction for English-Chinese CLIR*, in *Proceedings of the 2007 ACM symposium on Applied computing*. 2007, ACM: Seoul, Korea.

[22] Pirkola, A., et al., *Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings.* Information Retrieval, 2001. 4(3): p. 209-230.

[23] Lee, C.-J., J.S. Chang, and J.-S.R. Jang, *Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources.* ACM Transactions on Asian Language Information Processing (TALIP), 2006. 5(2): p. 121-145.

[24] Chinchor, N.A., *Overview of MUC-7/MET-2*, in *Proceedings of the 7th Message Understanding Conference (MUC-7)*. 1997.

[25] Anonymous. *Wikipedia*. 3 September 2009]; Available from: http://www.wikipedia.org.

[26] Nguyen, D., et al., *WikiTranslate: Query Translation for Cross-Lingual Information Retrieval Using Only Wikipedia*, in *Evaluating Systems for Multilingual and Multimodal Information Access*. 2009. p. 58-65.

[27] Gao, W., et al. *Cross-lingual query suggestion using query logs of different languages*. in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07*. 2007: ACM Press.

[28] Pourmahmoud, S. and M. Shamsfard. *Semantic Cross-lingual Information Retrieval*. in *2008 23rd International Symposium on Computer and Information Sciences, ISCIS 2008*. 2008.

[29] Shamsfard, M., A. Nematzadeh, and S. Motiee, *ORank: an ontology based system for ranking documents.* International Journal of Computer Science, 2006. 1(3): p. 225-231.

[30] Yuan, S. and S. Yu, *A new method for cross-language information retrieval by summing weights of graphs*, in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery*, J. Lei, Editor. 2007, IEEE Computer Society. p. 326 - 330.

[31] Wu, K. and B. Lu, *A refinement framework for cross-language text categorization*, in *Springer Lecture Notes in Computer Science*, H. Li, Editor. 2007, Springer-Verlag: Berlin Heidelberg. p. 401-411.