International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 4, Issue 3 (May-June, 2016), PP. 258-261

DEVELOPMENT OF A FRAMEWORK FOR QUERY TRANSLATION AND ITS AMBIGUITIES

Pratibha Bajpai[1], Parul Verma[2], S.Q.Abbas[3]
[1]Research Scholar, [2]Assitant Professor, [3]Professor
Department of Information Technology, Amity University, Lucknow, India
pratibhabajpai@gmail.com

Abstract— **Internet usage is increasing at a rapid rate. Anyone can access all sorts of information from the web at any time. But the language diversity of web pages restricts this access of information. Cross Lingual information Retrieval provides the solution to this problem, by allowing user to post their query in the native language and retrieve the documents in other languages. Machine readable dictionaries are the most economical resource of cross language translation. These dictionaries usually provide more than one translations in target language for source query terms. These translations need to be disambiguated to achieve the best translation for a query word. Once we achieve the correct translations, the cross lingual information system can match the monolingual information retrieval on performance grade. In this paper we propose a framework for query translation and disambiguation to achieve an efficient cross lingual information system.**

I. INTRODUCTION

Information Retrieval (IR) provides relevant documents to user depending on her information needs. This information need in technical terms is called, query. Query can be defined as the request expressed as search keys in a form that the retrieval system is able to process. With the advent of World Wide Web, IR systems have impact on every field like, entertainment, business, education etc. and answering every user's query somewhat effectively. With the explosion of information on web and that also in multitude of diverse languages, there is an urge for IR systems to cross language boundaries. Such systems will facilitate user's to retrieve documents in any language with query in one language.

CLIR helps in removing linguistic gap between the user query and documents retrieved. Cross language Information Retrieval can thus be defined as retrieving documents in language different from the language of request [1]. The query language is called Source language and language of documents is referred as Target language. Typically the source language happens to be native language of user and the target language can be a language in which the user can only read documents while typing query may be tough. To overcome the language disparity, CLIR engines are required to incorporate some features for language translation, if meaningful comparison is to be done between query representation and document representation. There are two general approaches to such translation. We can either translate the user query into the language of document collection or translate document collection into query language. With a CLIR system such globalized information and linguistic multiplicity will no longer be a barrier for accessing information across languages on the web.

The drive for evaluation of monolingual and cross-lingual retrieval systems started with Cross-Language Evaluation Forum (CLEF) in European languages and NTCIR in Chinese-Japanese-Korean languages. It is only in the recent past that the Indian languages have gained importance in evaluation. From 2008, a specific campaign focusing on Indian languages started with the Forum for Information Retrieval Evaluation (FIRE).

The paper organization is as follows: section 2 analyzes query translation process and translation ambiguity. Section 3 discusses various components of our proposed framework. Finally section 4 draws the conclusion.

II. ANALYSIS OF QUERY TRANSLATION PROCESS AND TRANSLATION AMBIGUITY

A. Query Translation or Document Translation

There are two types of translations available in CLIR. We can either translate the user query into the language of document collection or translate document collection into query language. Query translation is simple as much syntactic knowledge need not be considered as contrast to document translation which is time consuming, expensive and hard to implement. Also literature review shows no performance advantage of document translation over query translation [2].

B. Keyword Selection

Linguistics features in general are important in CLIR. Word inflection is a well known source of setback in IR. Word inflection is a process in which base form of a word is tailored to express various grammatical meanings for instance write takes the forms 'write', 'written' and 'wrote'. User expects IR system to retrieve all documents containing all the inflected forms against the user query word 'write'. This can be achieved by using a word Stemmer, which maps inflected word forms into a common base form both in documents and query.

Other possible improvement can be achieved in CLIR, by removal of Stop words. Stop words are words that are non-significant from linguistic point of view for instance 'if', 'an', 'the'. The process of removing stop words can be done by

using a stop word list that enumerates all words with little meaning.

C. Selection of translation resource

Another issue in CLIR is selecting translation resource. Depending on the resource, three different techniques exist in CLIR: Dictionary based CLIR, Corpora based CLIR and Machine translator CLIR.

Machine translation as the name suggests, uses software to translate text from one language to another, Dictionary based translation uses Machine Readable Dictionary (MRD) and Corpora based translation use parallel and comparable corpora to translate query. DB-CLIR is the least resource intensive CLIR technique but suffers from the problem of ambiguity. Machine translation overcomes this problem by returning only one translation but leading to loss of recall in document retrieval. Parallel corpora can provide more accurate translation knowledge but due to their scarcity, they are not a common source of translation for many language pairs. This leaves machine readable dictionaries as the most viable resource for CLIR.

D. Ambiguity Removal

Dictionary based CLIR suffers from lexical ambiguity problem. Ambiguity refers to the increase of irrelevant search key senses as there are a large number of words in natural languages which carry more than one meaning. For instance, word 'bank' has three senses. Different senses refer to a 'financial institution' or 'river bank' or 'reservoir'. Hull and Grefenstette define translation ambiguity as the difficulty of choosing the right translation for given query terms [3]. These ambiguities in search keys lead to unsuccessful retrieval of relevant documents.

The main problem with DB-CLIR is selecting the appropriate translation of query words, without much contextual information being available. According to Salton, CLIR can match monolingual IR on performance scale, if necessary translations are carried out accurately [4]. As a matter of fact, CLIR systems do not perform at that level.

III. Proposed Framework

Considering the above discussed issues, we propose a framework for effective Cross Lingual Information Retrieval.

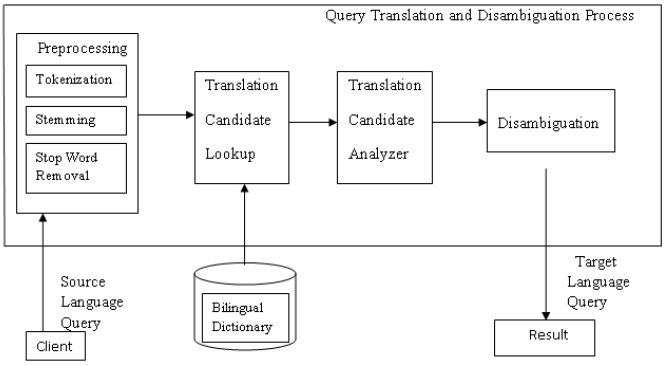

Figure 1. Proposed Framework

A. Preprocessing

The first step is preprocessing of query terms to speed up the translation process without affecting the retrieval quality. This preprocessing involves tokenization, stemming and stop word removal. It ends up with a bag of vital words.

3.1.1. Tokenization: By tokenization, we isolate those parts of a source query, which are significant as translation candidates for MRD. These candidates are referred as Tokens.

3.1.2. Stop Word Removal: Standard stop word list, which is easily available, is used to remove stop words from the query.

3.1.3. Stemming: Next step will be to map all the different inflected forms of a word to the same stem. For this purpose we use advanced stemming algorithm like Porter stemmer, Snowball stemmer etc.

B. Translation

A bilingual machine readable dictionary is used to find translation candidates for query search terms.

C. Analyzer

Dictionary translation leads to spurious equivalent translations in target language. All the translations are not desirable, depending on the context of the query. So to improve the computation speed, we need to analyze these dictionary translations to identify the desired ones.

Giang use Word Distribution algorithm. For each Vietnamese token extracted from query preprocessing, there exists a list of translation candidates in English. For each English word, Giang et al. count the number of times it appears in the training corpus and thus create a word distribution. The translation for a Vietnamese word is created by selecting the candidate with highest distribution value. Finally, the English query is created by joining selected translations. [5]

Other strategy is to take the first n translation candidates from dictionary for each source-language query term. Another strategy uses simple word by word translation by taking random nth translation equivalent from dictionary. The limit for random number is set by observing the number of meanings for a word in bilingual dictionary.

The other idea can be of using part-of-speech (POS) tags for translation candidates to select only translations having the same POS with that of the source query term. Davis and Ogden [6] applied a part-of-speech tagger to English queries. Spanish translation was selected as a search term, only if the POS tag of a Spanish equivalent listed in an English–Spanish dictionary matched with that of the English query term.

D. Disambiguation

Very often, most of the translation candidates found in bilingual dictionaries are irrelevant to the semantic meaning of query, thereby making disambiguation crucial for DB-CLIR.

Researchers have reported several approaches [7, 8, 9, 10, 11, 12, 16, 17] to resolve query translation ambiguity in DB-CLIR. The easiest one is to make use of all translation candidates provided by dictionary for each query word with equal weight [8]. This can be treated as no sense disambiguation. Other approach for disambiguation is by computing coherence score of a translation candidate measured

using co-occurrence statistics to the entire query. A translation candidate is assigned high coherence score when it co-occurs frequently with the translation of other query words. [8, 9, 11, 13, 18] use the translation candidates with highest coherence score while in [7, 14] multiple translations are selected provided their coherence score exceeds predefined threshold.

E. Evaluation of result obtained

The effectiveness of the proposed framework will be measured by two standard variables: Recall and Precision. The variables are defined as

$$\text{Recall} = \frac{\text{number of relevant documents retrieved}}{\text{number of relevant documents in the collection}}$$

$$\text{Precision} = \frac{\text{number of relevant documents retrievec}}{\text{total number of documents retrieved}}$$

A system capable of retrieving all relevant documents have high recall and one in which most of the retrieved documents is relevant has high precision. These measures evaluate the quality of unordered set of retrieved documents.

To measure ranked lists, precision and recall can be considered in combinations, for example, precision can be plotted against recall after each retrieved document. Average precision is the average of the precision value obtained at each relevant document is retrieved. It rewards systems that rank relevant documents high. An example from [15] exemplifying average precision is the following: a query with four relevant documents retrieved at ranks 1, 2, 4 and 7. The actual precision after each relevant document is retrieved is 1, 1, 0.75 and 0.57. Computing the mean, which is 0.83, gives the average precision over all relevant documents for this query as 0.83. Precision can also be measured at standard recall levels (0 to 1 in increments of 0.1).

IV. CONCLUSION

Major source of problem with dictionary based CLIR is sense ambiguity as dictionaries offer a good number of senses for natural language words. The proposed framework removes this ambiguity in two steps. In analyzer phase we will try to reduce target language translations to a few and then in disambiguation phase we will find the correct sense of query word. The proposed framework aims at finding the correct translation of source language words and thus meets the effectiveness of monolingual information retrieval.

REFERENCES

1. P. Iswarya and V. Radha, 2012. “Cross Language Text Retrieval: A Review”. International Journal Of Engineering Research And Applications. 2(5), pp.1036-1043.
2. Donnla Nic Gearailt, 2005. Dictionary characteristics in cross-language information retrieval.Technical Report (Number 616), University of Cambridge.
3. David A. Hull and Gregory Grefenstette 1996, Querying across languages: A dictionary-based approach to multilingual information retrieval. In Proceedings of the 19th International Conference on Research and Development in Information Retrieval, pages 49–57.
4. G. Salton. Experiments in multi-lingual information retrieval. Information Processing Letters, 2(1):6{11, 1973. also Technical Report TR 72-154 at Cornell University, 1972.
5. L. T. Giang, V. T. Hung and H. C. Phap. Experiments With Query Translation And Re-ranking Methods In Vietnamese-English Bilingual Information Retrieval. SOICT’13, December 05 - 06 2013, Danang, Vietnam.
6. Davis, M. & Ogden, W. C. (1997). QUILT: Implementing a Large-Scale Cross-Language Text Retrieval System. In Proceedings of the 20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, USA, pp. 92-98.
7. Jang, M.-G.; Myaeng, S. H.; and Park, S. Y. 1999. Using mutual information to resolve query translation ambiguities and query term weighting. In *ACL ’99*.
8. Kraaij, W., and Pohlmann, R. 2001. Different approaches to cross language information retrieval. In Daelemans, W.; Sima’an, K.; Veenstra, J.; and Zavrel, J., eds., *Computational Linguistics in the Netherlands 2000*, number 37 in Language and Computers: Studies in Practical Linguistics, 97–111. Amsterdam: Rodopi
9. Adriani, M. 2000a. Dictionary-based CLIR for the CLEF multilingual track. In *CLEF ’00*.
10. Adriani, M. 2000b. Using statistical term similarity for sense disambiguation in cross-language information retrieval. *Inf. Retr.* 2(1):71–82.
11. Gao, J.; Nie, J.-Y.; Xun, E.; Zhang, J.; Zhou, M.; and Huang, C. 2001. Improving query translation for crosslanguage information retrieval using statistical models. In *SIGIR ’01*, 96–104. ACM Press.
12. Gao, J.; Zhou, M.; Nie, J.-Y.; He, H.; and Chen, W. 2002. Resolving query translation ambiguity using a decaying cooccurrence model and syntactic dependence relations. In *SIGIR ’02*, 183–190. ACM Press.
13. Davis, M. W. 1996. New experiments in cross-language text retrieval at NMSU’s computing research lab. In Harman, D. K., ed., *TREC-5*. NIST.
14. Maeda, A.; Sadat, F.; Yoshikawa, M.; and Uemura, S. 2000. Query term disambiguation for web cross-language information retrieval using a search engine. In *IRAL ’00*, 25–32. ACM Press.
15. Peters, C. ed. 2001. Cross-Language Information Retrieval and Evaluation. Workshop of the Cross-Language Evaluation Forum, CLEF 2000 Lisbon, Portugal, September 2000. Revised Papers. Lecture Notes in Computer Science 2069. Berlin: Springer.
16. L. T. Giang, V. T. Hung and H. C. Phap, 2013. “Experiments with Query Translation And Re-ranking Methods In Vietnamese-English Bilingual Information Retrieval”. SOICT’13, Danang, Vietnam, December 05 – 06.

17. A. Duque, L. Araujo, and R. Juan, 2015. "CO-graph: A new graph-based technique for cross-lingual word sense disambiguation" in Natural Language Engineering, Volume 21, Special Issue 05, pp 743-772.
18. S. Varshney and J. Bajpai, 2013. "Improving performance of English-Hindi cross language information retrieval using transliteration of query terms" in International Journal on Natural Language Computing (IJNLC) Vol. 2, No.6.