International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 319-321
e-ISSN:2278-621X

CROSS LANGUAGE QUERY PASSING AND INFORMATION RETRIVAL

Aidan Menezes[1] & Royston Fernandes[2]

Abstract- Machine Multilingual information is overflowing on internet these days. This increasing diversity of web pages in almost every popular language in the world should enable the user to access information in any language of his choice. But sometimes it is difficult for a user to write her request in a language which she could easily read and understand. This makes cross-language information retrieval (CLIR) and for Web applications a valuable need of the day. It increases the accessibility of web users to retrieve information in any language while post their queries in their native language.
Keywords– Cross lingual Information Retrieval, Query Translation.

1. INTRODUCTION

A cross language information retrieval (CLIR) system is a system for retrieving documents across language boundaries. A query written in one language should be translated into a representation for finding documents in another language. In this paper, we propose a method for the translation, which uses a parallel corpus and can readily be combined with the information mapping approach

A classic IR system accepts the user information need in a form of query and gives back the documents that are relevant to the user need. With the explosion of knowledge on the web, it became necessary to break the language barriers for the monolingual IR systems. This may allow the users of IR systems to give query in one language and retrieve documents in different languages.

IR system, with different source and target language is called CLIR system. Cross-Lingual Information Retrieval (CLIR) translates the user query (given in source language) into the target language, and uses translated query to retrieve the target language documents. 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). This resulted in the development of large document collection in some Indian languages like Bangla, Hindi, Marathi and Tamil.

2. DIFFERENT TECHNIQUES FOR CLIR

Based on different translation resources, three different techniques have been identified in CLIR: Dictionary based CLIR, Corpora based CLIR and Machine translator based CLIR.

2.1 Machine Translation

Machine translation, in simple terms, is a technique that makes use of software that translates text from one language to another language. But machine translation is not all about substitution of words from one language to another only; rather it also involves finding phrases and its counterparts in target language to produce good quality translations. Machine translation is of three types:

2.2 Rule Based Machine Translation

Multilayer perceptron(MLP) consists of atleast three layer of nodes. MLP is a class of feed forward artificial neural network. Back propagation technique which is a supervised learning technique is utilized for training. MLP are sometimes referred as "vanilla" neural networks mostly when they have a single hidden layer. The network is a network of simple processing elements which work together to produce a complex output. A multilayer feed forward network has input layer, one or more hidden layer and an output layer.

2.3 Statistical Machine Translation

Statistical machine translation generates translations using statistical methods based on bilingual text corpora. Dan Wu & Daqing He (2010) conducted a series of CLIR experiments using Google Translate for translating queries. Their results show that with the help of relevance feedback, MT can achieve significant improvement over the monolingual baseline, no matter

[1] MCA V Semester, St Aloysius College, AIMIT, Mangalore, Karnataka, India
[2] MCA V Semester, St Aloysius College, AIMIT, Mangalore, Karnataka, India

whether the query length are short or long. Kraaij & Simard(2003) experimentally claim that web can be used for automatic construction of parallel corpus which can then be used to train statistical translation models automatically.

2.4 Example Based Machine Translation

Example based MT reads similar examples in the form of source text and its translation from the set of examples, adapting the examples to translate a new input. Sato and Nagao (1990) investigated the problem of example selection by approximate matching of input sentences and example sentences, using a similarity measure based on the syntactic similarity of dependency tree structures of a sentence pair in question and on the word distance of corresponding words, which were predefined in a thesaurus. Sumita et al. (1990) looked into example-based translation of Japanese noun phrases of the pattern [N1 no N2] into English as [N2 prep N1] or [N1 N2], based on a distance measure for the input phrase and example phrase, calculated as a linear weighted sum of the distances of the three sub-parts, each of which is predefined in a thesaurus.

2.5 Corpus Based Cross Lingual Information Retrieval

Corpus based CLIR methods use multilingual terminology derived from parallel or comparable corpora for query translation and expansion. There are two types of corpus

2.6 Parallel Corpus

A parallel corpus is a collection where texts in one language are aligned with their translations in another language. Several systems have been developed to mine large parallel corpora from the web. Wang and Lin (2010) give a method which first identifies a set of seed URLs and crawl candidate bilingual websites. The obtained pages are cleaned and bilingual texts collected to construct comparable corpora. Wang et. al. (2004) exploit the bilingual search result pages obtained from a real search engine as a corpus for automatic translation of unknown query terms not included in the dictionary. They propose a PAT-tree based local maxima method for effective extraction of translation candidates. The approach gives excellent results.

2.7 Comparable Corpus

Comparable corpus, on the other hand, consist of texts that are not translations, but share similar topics. They can be, e.g., newspaper collections written in the same time period in different countries. Sadat Fatiha (2011) exploit the idea of using multilingual based encyclopedias such as Wikipedia to extract terms and their translations to construct a bilingual ontology or enhance the coverage of existing ontologies. The method show promising results for any pair of languages. Qian & Meng (2008) expanded Chinese OOV phrase with its partial English translation and submitted to the search engine. The translation of OOV words is mined by preprocessing the snippets obtained to extract the main text from the web page. The strings obtained are sorted by weighted frequency to output the top n translation of OOV phrase. The method proves to obtain the translation with high time efficiency and high precision.

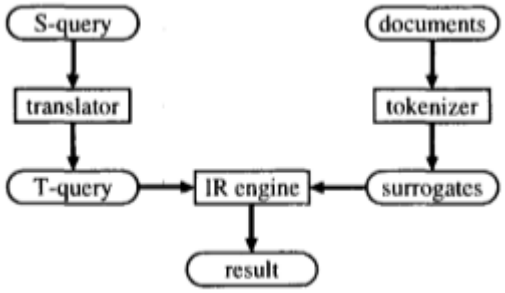

Figure 1 – The overall design of CLIR system.

3. DIFFERENT PHASES OF CLIR

3.1 Preprocessing

The first step in any CLIR system is preprocessing of query terms to speed up the translation process without affecting the retrieval quality. This preprocessing is done using tokenization, stemming and stop word removal.

3.2 Tokenization

Tokenization is defined as an attempt to recognize the boundaries between words and isolate those parts of a query which should be translated in the source query.

3.3 Stemming

It maps all the different inflected forms of a word to the same stem. For languages like English which have weaker inflections, simple stemming algorithms can be used. Such algorithms only remove plural endings. In languages with stronger inflections,

suffices are joined to the stem end to end. The advanced stemming algorithm can recognize such multiple endings and remove them in an iterative fashion. Porter stemmer, Snowball stemmer etc. are well known advanced stemming algorithm.

3.4 Query Translation

In Query Translation, the given query is converted from Source language to Target language and the obtained query searches the database to get the documents in Target language. Query Translation often suffers from the problem of translation ambiguity and this problem is amplified due to the limited amount of context in short queries. Query translation can be done using any one technique including machine translation, dictionary based or corpus based method. The techniques have already been discussed in section 2. The query translation is quiet complex while translating English to Hindi query as the two languages are morphologically different from each other.

4. CONCLUSION

The respective work with regard to Indian languages has gained impetus in last decade and there is much to be explored in this field. It is quite obvious from the observations that there is still a scope of improvement in the performance level of CLIR. We presume that the proposed prototype system will prove to be competent with other existing systems.

5. REFERENCES

[1] Ballesteros, L, and Bruce W Croft, "Phrasal Translation and Query Expansion Techniques for Cross Language Information Retrieval". In: Proceedings of 20th International ACM SIGIR Conference in Research and Development in IR 1997.

[2] Ballesteros, L., and Croft, W.B. 1998. "Resolving ambiguity for cross-language retrieval." In Proceedings of SIGIR Conference, pages 64-71, 1998.

[3] Chawre, S. M., Srikantha Rao. Domain Specific Information Retrieval in Multilingual Environment, International Journal of Recent Trends in Engineering, 2, 4, 179-181, 2009.

[4] Chinnakotla Kumar Manoj, Ranadive Sagar, Bhattacharyya Pushpak and Damani P. Om "Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007", in the working notes of CLEF 2007.

[5] David A. Hull and Gregory Grefenstette. 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, 1996.

[6] Dr. Saraswathi, S., Asma Siddhiqaa, M., Kalaimagal, K., and Kalaiyarasi M. BiLingual Information Retrieval System for English and Tamil, Journal Of Computing, 2,4, 85-89, April 2010.

[7] Grefenstette, G. (1998b). The problem of cross-language information retrieval. In Grefenstette (1998a), pages 1-9.

[8] Hsu Hung Ming, Tsai Feng Ming, and Hsin-Hsi Chen Query Expansion with ConceptNet and WordNet: An Intrinsic Comparison. In : AIRS 2006, LNCS 4182, (2006) 1-13.

[9] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. Query Expansion by Mining User Logs IEEE. Transactions on Knowledge and Data Engineering, Vol. 15(4) 2003.

[10] Flournoy, R., Masuichi, H. and Peters, S. (1998). "Cross-Language Information Retrieval: Some Methods and Tools". In Proceedings of 14th Twente Workshop on Language Technology.

[11] Schütze, H. (1995). "Ambiguity Resolution in Language Learning: Computational and Cognitive Models". PhD thesis, Stanford University, Department of Linguistics.

[12] Salton, G. and MacGill, M. (1983). "Introduction to Modern Information Retrieval". McGrowHill.