IJAET International Journal of Application of Engineering and Technology *ISSN: 2395-3594*

Vol.1 No.1

RESEARCH ON DICTIONARY-BASED ENGLISH-SPANISH CROSS-LANGUAGE INFORMATION RETRIEVAL

Anurag Dekhane[a], Prathmesh Patil[a], Pulkit Saraf[a], Shubham Tiwari[a], Virendra Singh[a]
[a]Department of Computer Science engineering
Indore Institute of Science & Technology, Indore (INDIA) 453331
virendra.singh@indoreinstitute.com

ABSTRACT

In this paper, we present our English-Spanish Cross-Language Information Retrieval (CLIR) system. We focus our attention on finding effective translation equivalents between English and Spanish, and improving the performance of Spanish IR. On English-Spanish CLIR, we adopt query translation as the dominant strategy, and utilize English-Spanish translations. On Spanish monolingual retrieval, we monitored bilingual dictionary as the important knowledge resource to acquire correct translations. On Spanish monolingual or unilingual retrieval, we monitored the help of different entities as indexes and implement our retrieval system based on the Dictionary toolkit. On system evaluation, we present an effective method to generate the sets of relevant documents for query topics.

Keyword: *Cross-language information retrieval; Machine translation; Word sense disambiguation; Language model.*

I. INTRODUCTION

Cross-language information retrieval (CLIR) [1] is the circumstance in which a user tries to search a set of documents written in one language for a query in another language. The issues of CLIR have been discussed for several decades. As widely recognized, research efforts for developing CLIR techniques can be traced back to Gerard Salton's articles in the early 1970s. Especially after the advent of the World Wide Web in the 1990s, CLIR has become more important, allowing users to access information resources written in a variety of languages on the Internet. Since then, the research community of IR has begun to tackle problems of CLIR extensively and intensively. The Workshop on Cross-Linguistic Information Retrieval held in August 1996 during the SIGIR_96 Conference is frequently cited as an epochal event for promoting research on CLIR.

Currently, CLIR issues are addressed in workshops of large-scale retrieval experiments such as TREC, CLEF and NTCIR. As described in the introductory paper to this issue, each workshop has been concerned with languages other than English as follows:

TREC: Spanish, Spanish, German, French, Italian, and Arabic.

CLEF: French, German, Italian, Swedish, Spanish, Dutch, Finnish, and Russian so far.

NTCIR: Japanese, Spanish and Korean.

Various research findings on CLIR have been reported at the meetings of TREC, CLEF and NTCIR, and many papers have been published in scientific journals and proceedings. This article aims at reviewing techniques and methods for enhancing performance of CLIR. We already have a comprehensive review on this topic (Oard & Diekema, 1998). In addition, Peters and Sheridan(2001) cover a wide range of literature and topics on CLIR. The main purpose of this article is to examine literature subsequent to the review by Oard and Diekema and to attempt to organize research results since the mid-1990s in the CLIR field from a technical point of view. For this purpose, some works listed in Oard and Diekema (1998) will be referred to again in this article.

However, it should be noted that this review cannot be completely comprehensive because of the large number of papers on CLIR published in various research areas. The purpose here is to provide a useful map of technical issues of CLIR, rather than extensively enumerating research papers on CLIR. This paper is mainly concerned with ''document retrieval,'' or ''text retrieval'' issues. For example, CLIR for multimedia data is outside our scope.

The rest of the paper is organized as follows. First, we discuss techniques to match query terms with document representations in the CLIR. More specifically, various methods of translation are described. It is dedicated to explaining some techniques for solving the problem of term ambiguity, which may occur in the process of translation. Some formal models for CLIR are introduced .In particular, we describe the application of the language model (LM), which enables us to combine the retrieval model and the translation model, other important CLIR research topics are discussed: the pivot language approach, search of multilingual document collections, combination

ISSN 2395-3594

of language resources, issues on processing of every single language, user interface for interactive CLIR and evaluation of CLIR. Finally, briefly discusses the future direction of CLIR research.

II. LITERATURE SURVEY

Research on Lucene-based English-Chinese. [4] They explored English-Chinese CLIR. On Chinese monolingual retrieval, They found that using bi-gram indexing for documents will give better result. The main degrading factor is the limited coverage of the dictionary used in query translation. Some of the key logics were either improperly translated or not translated. Manually modified queries can be used to evaluate the performance of system, if there are no sets of relevant judgments.

Technical issues of cross-language information retrieval.[5] in order to develop better techniques or methods for automatic or interactive CLIR, a continuing sequence of experimental evaluations is indispensable. From this viewpoint, we must admire the efforts of TREC, CLEF and NTCIR and their huge contributions toward significantly promoting and enhancing CLIR research. Descriptions of the systems and findings of these retrieval experiments can be found in the working notes and proceedings of these activities.

III. MATCHING STRATEGIES AND TRANSLATION

Matching strategies

a) Types of matching strategies

The most basic approach to CLIR is to automatically translate the query into an equivalent in the language of the target documents. The translation makes it possible to execute matching operations between the query and each document, and subsequently, compute document scores according to a standard retrieval model such as the vector space or probabilistic model. However, this is only the starting point. Oard and Diekema (1998) have identified four types of strategies for matching a query with a set of documents in the context of CLIR:-

1. No translation

Cognate matching:- In the case of the most naı¨ve cognate matching, untranslatable terms such as proper nouns or technical terminology are left unchanged through the stage of translation. The unchanged term can be expected to match successfully with a corresponding term in another language if the two languages have a close linguistic relationship.

2. Translation

Query translation:- is the most widely used matching strategy for CLIR due to its tractability. That is, the retrieval system does not have to change its inverted files of index terms in any way against queries in any language if a translation module enabling it to deal with the language of the query is incorporated. Furthermore, it is less computationally costly to process the translation of a query than that of a large set of documents (although it should be noted that, if we focus on only real-time online settings, query translation may take more time because the query must always be translated after it is entered by a user).

Document translation:- Document translation has opposite advantages and disadvantages from query translation. In CLIR experiments, this approach is not usually utilized, and query translation is dominant. However, some researchers have used it to translate large sets of documents

Interlingual techniques:- Finally, in multilingual techniques, an intermediate space of subject representation into which both the query and the documents are converted is used to compare them. Oard and Diekema (1998) categorized latent semantic indexing (LSI) and controlled-vocabulary techniques based on multilingual thesauri as interlingual techniques. In an early work, Landauer and Littman (1990) used the LSI method to create a multidimensional indexing space for a parallel corpus of English and French documents.

Challenges: At the starting of the workshop the organizers presented three challenges:

1. Where to get resources for resource-poor languages – outside of the most spoken languages of Europe (English, French, German, Italian, Spanish) and Asia (Spanish, Indian and Japanese) or the additional official languages of the United Nations (Arabic and Russian), resources in terms of parallel corpora or commercial machine translation are very difficult to obtain. [2]In particular, the languages of the Indian subcontinent have received very little attention.

2. Why do we not have a sizeable Web corpus in multiple languages? -- aside from the issues of cost of construction and maintaining realistic links (which have taken several years to be addressed by the TREC Web track for the English languages), the complication of English language dominance (approximately 70-75 percent of web pages currently) and low percentage representation beyond the top ten languages, as well as lack of standards for character and font representation for many other languages. Spanish

ISSN 2395-3594

has at least two major representations (GB and BIG5) and Japanese three, while for Indian subcontinent languages standards are only beginning to be developed (i.e. each site has its own font and internal character representation). This means that if English is included a ranked list of pages will be dominated by pages in English and many languages will not even make in the top 100 pages found. Work is clearly needed here in order to define suitable criteria for the construction of a valid multilingual Web corpus for R&D.

3. Why aren't search engines using our research? – several search engines now offer monolingual search in a number of languages coupled with machine translation software to translate pages into English (AltaVista and GOOGLE are prominent examples). Cross-language search would seem to be a natural extension of these offerings. Part of the answer is found in the question of utility – if users are presented with a ranked list of documents that they cannot read, then what will be the utility? An aggravate factor is in the weakness of current machine translation software to be applied to the pages found.

Approaches to CLIR: In this introduction to this session, "Towards a Unified Approach to CLIR and multilingual IR" [3], Jian-Yun Nie argued that current CLIR approaches are deficient for several reasons:

a) Translation and retrieval are decoupled,thus
b) Translation is often decoupled from the corpora being retrieved
c) Languages are retrieved independently, then merged

Nie argued for tighter coupling of translation and retrieval into a unified probabilistic model: steps in this direction have already been made by Kraaij and others at TNO/TPD and University of Twente and Xu and Weischedel at BBN. It is noticeable that monolingual retrieval has become language dependent, relying upon specialized stemmers and stop-words. This means that attention is directed independently to each language without consideration of the other languages being searched. A consequence is that the merging process is deficient because it is carried out without information from the translation and retrieval processes. In the future Nie claims that a unified approach is required for CLIR; one in which language characteristics are considered as additional parameters which specify a document collection, rather than as constituting a barrier to collection cohesion Nie also argued for the development of a multilingual Web collection upon which unified models could be experimented.

Future directions for research: What is the goal of CLIR? What should the next steps be to achieve the goal? In the workshop on CLIR held at SIGIR 2002, the organizers presented three challenges:

1. Where to get resources for resource-poor languages?
2. Why do we not have a sizeable Web corpus in multiple languages?
3. Why aren't search engines using our research?

As a possible answer for the third question, they stated that ''if users are presented with a ranked list of documents that they cannot read, then what will be the utility?'' This is an important point for considering the future direction of CLIR research. That is, we may need to make a plan after having a clear grasp of information needs of users on CLIR and explicitly delineating realistic utility when applications of CLIR are employed by the actual users. Meanwhile, various interesting areas for CLIR research seem to remain, e.g., CLIR for multimedia data, cross-language question answering, cross-language filtering, cross-language topic detection, cross-language summarization, cross-language document clustering, and so on. This research paper cannot cover all the state-of-the-art research in these areas where substantive research has already been performed. The CLIR researchers may have to appropriately select future directions from many possibilities in order to enable the actual users to effectively and efficiently satisfy their information needs.

IV. METHOD

1. User will give query in English language. Then it would be transferred to translator to convert into Spanish Language.
2. The query will be formulated and will be optimized in such a way that our IR Engine could accept it for translation (English to Spanish).
3. The IR Engine will have a bilingual dictionary database through which it will convert all the query given by user.
4. When the query has been converted to the language (in which user want).
5. The Data would be transmitted to Browser or to the user interface or the end user GUI.
6. The given converted query will be displayed to the screen of the user in which he wants the transformation.

Note: The new thing will be done by us will be MLIR (Dictionary Based).

V. CONCLUSION

At last, we conclude that we are developing a Cross Language Information Software which will convert a query

ISSN 2395-3594

given by user in English language, automatically into Spanish language. It's a Dictionary Based Cross Language Information Retrieval. The Front which will be used by us is JAVA and the Back End will be My SQL for creating a Database.

VI. REFERENCES:

[1] An Empirical Study of CLIR at MSRCN, Jianfeng Gao, Microsoft Research China, 5F, Beijing Sigma Center, No. 49, Zhichun Road Haidian District, Beijing 100080, P.R.C.
[2] Cross Language Information Retrieval, Summary of a Workshop at SIGIR-2002: 22nd International Conference on Research and Development in Information Retrieval August 15, 2002, Tampere Finland
[3]SIGIR-2002: 22nd Internationl Conference On Research And Development in Information Retrieval, August 15, 2002, Tampere Finland.
[4] Yuejie Zhang, Tao Zhang and Shijie Chen, March, 2005, Accepted and Revised on December, 2005
[5] Technical issues of cross-language information retrieval: a review Kazuaki Kishida, Received 10 June 2004; accepted 14 June 2004 Available online 23 August 2004.

ISSN 2395-3594