International Journal of Information Science and Management

Applying Natural Language Processing Techniques for Effective Persian- English Cross-Language Information Retrieval

H. Alizadeh, Ph.D.
Regional Information Center for
Science & Technology, I. R. of Iran
email: halizadeh@gmail.com

R. Fattahi, Ph.D.
Ferdowsi University of Mashhad
I. R. of Iran
email: fattahirahmat@gmail.com

M. R. Davarpanah, Ph. D.
Ferdowsi University of Mashhad, I. R. of Iran
email: mdavarpanah@gmail.com

Abstract

Much attention has recently been paid to natural language processing in information storage and retrieval. This paper describes how the application of natural language processing (*NLP*) techniques can enhance cross-language information retrieval (*CLIR*). Using a semi-experimental technique, we took Farsi queries to retrieve relevant documents in English. For translating Persian queries, we used a bilingual machine-readable dictionary. *NLP* techniques such as tokenization, morphological analysis and part of speech tagging were used in pre-and- post translation phases. Results showed that applying *NLP* techniques yields more effective *CLIR* performance.

Keywords: Cross-Language Information Retrieval, Natural Language Processing, Machine- Readable Dictionary, Evaluation.

Introduction

The vast amount of multilingual information on the Internet and other major information providers such as integrated databases implies that there is a crucial need for novel scholarly information retrieval systems. All information systems are in need of overcoming language barriers and helping users to find information needs in any foreign language. Finding relevant information in languages other than one's native one language is very important today. Access to all relevant documents needs more powerful and sophisticated retrieval systems. The ability of modern information retrieval systems to return the most relevant documents for a specific query has become more and more important in the age of extremely large collections, such as the World Wide Web. Although the linguistic diversity on the Internet seems to be useful at first sight, it can however prevent access to the needed information (Alizadeh, 2004). Today the task of information

retrieval is not restricted to traditional processes, but the larger goal, namely to overcome language barriers during the search and retrieval of information, must be achieved. Edwards (1994) estimated that there are more than 4500 living languages, thirty of which are used by at least 30 million people (Edwards, 1994). This implies that, to exchange information in a multilingual information society, one cannot be limited to a specific language. Internet, as a meeting place of these languages, has a multilingual nature. Statistics show that the use of the Internet in recent years has had significant growth. This growth rate, especially in the Middle East, South America and Africa is very significant. This geographical diversity also is associated with linguistic diversity; therefore, with the growth of the Internet resources in different languages, linguistic problems of searching for and retrieval of these resources have also increased.

Cross-language information retrieval (*CLIR*) is a good solution to overcome problems associated with language barriers. Cross-language information retrieval is a kind of information retrieval in which the query language is different from the document language. In *CLIR* system a user is not restricted to his own language, so he can formulate his query in his native language but the system returns documents in another language. This process will be carried out by translating the user's query into documents' language. Research in the area of cross-language information retrieval (CLIR) has focused mainly on methods for translating queries (Ballesteros& Croft, 1998).

Persian-English CLIR means the retrieval of documents based on queries formulated by a user in the Persian language, and the documents are in the English language. In other words, CLIR integrates the language of the searcher to the language of documents retrieved.

CLIR system simplifies the search process for multilingual users and enables those who know only one language to provide queries in their language and then get help from translators for using other languages' retrieved documents. With the increasing availability of machine-readable bilingual dictionaries, dictionary-based query translation has become a viable approach to Cross-Language Information Retrieval (Adriani, 2000).

Translation ambiguity happens in query translation because of the different nature of two natural languages involved in *CLIR*. To resolve this ambiguity, natural language processing (NLP) techniques deal with semantics of document texts and consider them as collection of meanings.

Problem Statement

At present, no Persian *CLIR* system is available to satisfy users' "cross-language information seeking behavior" needs. Due to lack of examining Persian language capabilities in this field, there is no information about the potential capability of the Persian language that could be applied to the processes of CLIR. Although in the last two decades

CLIR systems have been utilized in languages such as English, Spanish and French, as yet little is known about problems pertaining to the translation of queries in such systems in Persian. It is not clear that the use of natural language processing techniques, such as tokenization, morphological analysis and part-of-speech tagging w*ould yield results that would help clarify the unknowns that exist in the relationship between NLP techniques and the Persian language*. Accordingly, this research is a first attempt trying to find answers to the problems in question.

NLP and CLIR

Although natural language processing and information retrieval are two separate fields, effectiveness of using *NLP* techniques in *IR* has already been investigated. Among those who proposed the application of linguistic theories in IR are Sheridan and Seaton (1992) whose work showed the effectiveness of using linguistic techniques in processing natural texts.

The application of linguistic principles to the processing of natural texts has made available certain tools known as NLP agents. *NLP* is the analysis of natural language texts for the purposes of IR, machine translation, text generation and so on. Of the different levels of *NLP*, the two levels of morphology and syntax are widely used in *CLIR*. In this paper using these levels we examine the possible impact of applying NLP techniques on CLIR system and try to investigate its effectiveness.

Morphological analysis

Morphology is a linguistic concept which deals with internal structure of words (Lyons, 1981). From the morphological view, a word is a kind of lexeme which may take different forms called inflected forms. Persian morphology is based on affixes, mostly suffixes, and some prefixes. For example, the Persian verb "raft) ” went) appears in some other forms like "miraft" , "rafti", "raftam". These are inflected forms of the word "raft".

Since all the inflected forms of a word are not included in a dictionary, query translation process faces some problems in *CLIR*. In the process of *CLIR* query translation, some words are not translated by electronic dictionaries, so they must be omitted from target queries which results in a poor retrieval. By using morphological analysis technique, the internal structure of a word is obtained making it possible to recognize the base form of a word and its affixes. Removing affixes is called normalization which can help with the translation of search queries by dictionaries. Those words which are not translated by the dictionary (because of the presence of affixes) are normalized by removing the affixes and sending them back for translation. Researchers like Porter (1980) and Hedlund (2003), approved the advantages of affix removal in IR process.

Among not translated words are out of vocabulary words (OOV) which can be proper

names, technical terms and loan words. Out of vocabulary words are not translated even after morphological analysis. This type of words can be transliterated using the target language alphabet and be added to final queries.

Syntactic level

Syntactic level of NLP techniques has several applications in IR. One of these applications is "tokenization". By tokenization, the words and other items in a sentence are recognized. Tokenization is the f*ir*st step in applying *NLP* techniques. It can be done in different levels such as sentence and word. With the use of tokenization the boundaries between words are recognized and those parts of a query which should be translated are identified. Some components of queries such as punctuation signs, dates and abbreviations will be detected by tokenization and will be omitted before translation.

"Part-of-speech tagging" is another useful syntactic analysis which can be used in *CLIR*. Phrase detection and translation are the most difficult task in *CLIR*. Problems of phrase identification and translation are already discussed in many researches. Queries are usually made of words and phrases. In most cases the meaning of a multi-word phrase is different from the total meanings of its constituent words. So a word by word translation of phrases results in retrieval of *ir*relevant documents. *NLP* suggests part of speech tagging for solving this problem. By assigning syntactic labels to each word in a query, and with regards to the structure pattern of each language, it is possible to recognize phrases and then translate them as units.

For example, patterns like 'noun noun' imply a noun phrase, and then a string like " سازمان ملل " (the United Nations) is a good candidate of being a phrase. In this research, we examined efficiency of part of speech tagging and phrasal translation on Persian *CLIR* performance.

Review of Literature

Literature related to Persian CLIR is scarce. Few works like Davarpanah (2009) who presented an aggregated methodology for construction of the stop word list in Persian language and generated a generic Persian stop word list, and Mehrad and Naseri (2008) who published a work in the field of NLP and IR, can be mentioned. A dictionary based experiment in French-English *CLIR* showed that word-by-word translation can decrease *CLIR* effectiveness by %40 to %60 compared with monolingual retrieval (Hull& Grefenstette, 1996). When the same researchers repeated their research by using phrasal translation, the *CLIR* effectiveness improved up to %91 of monolingual retrieval. Ballesteros and Croft (1998) in another work on Spanish-English *CLIR* achieved similar results. They indicated that lack of phrase coverage in a dictionary was not conducive to phrase translation. They believed that translating multi-term concepts as phrases was an

important step in reducing translation error. In their experiment, they compared the advantages of using a phrase dictionary with that of the co-occurrence method to translate phrases. They then used co-occurrence (CO) statistics to reduce ambiguity by inferring the correct translation of phrases not translatable via their phrase dictionary and compared the effectiveness of the two methods through a word-by-word translation as a baseline.

Chen (2002) has also used statistical method for identifying phrases in Chinese-English *CLIR*. His findings showed that phrasal translation in comparison to word-by-word translation increased retrieval effectiveness. He also emphasized using lexical sources with a good coverage of phrases.

Problems with inflected words which are not translated in *CLIR* process are examined in other languages. Hedlund(2003) in Finnish-English *CLIR* used stemming and morphological analysis to solve the problem of untranslated words. His findings showed that normalizing inflected words results in their translation which can improve effectiveness of *CLIR* processes. Other researchers like Porter (1980) have already justified stemming usefulness for IR.

Methodology

To investigate the effectiveness of applying *NLP* techniques to Persian-English *CLIR*, we examined different retrieval approaches. To do this, we used 40 TREC English queries which were first translated into Persian by human translators to obtain our Persian (original) set of queries and then translated them back to English using Farsidic[1] online dictionary. This is a preferred method in dictionary-based *CLIR* research (Pirkola, 2001). Aljlayl and Phir (2001) also re-iterate "This method is often used in dictionary based CLIR studies".

Our translation resource was Farsidic. Farsidic is a bilingual Internet dictionary which is chosen because it is free and available to use online: also it is a general dictionary and compatible with our query set with general domains. It provides most common translations first and suits the first match method used in this research.

Each CLIR query has three fields: 1- title 2- description (which describes information need) 3- narrative (relevance criteria).

We used First match method for translating query terms. Dictionaries usually provide several equivalents for each word some of which are not proper translations of the word. Choosing wrong translation results in translation ambiguity and causes false drop in retrieval results. Some dictionaries (such as Farsidic dictionary) provide the most common translation as first match. Using first match in such dictionaries decreases translation ambiguity.

We used *NLP* techniques in pre and post-translation stages of *CLIR*. *NLP* techniques used in this research are tokenization, Morphological Analysis and part-of-speech tagging. In the time of conducting this research, no suitable Farsi *NLP* tool was available, so the

processes were carried out manually.

Pre-translation *NLP* techniques used in this research (tokenization and stop word removal) helped us recognize those parts of Persian queries which are to be translated. Below is an example of tokenizing a Persian query استفاده از نیروی باد.(using the wind energy) before translation:

استفاده/token/نیروی/token/باد /token

As it can be seen, we first omitted the word از [i.e., from] because it is a preposition and belongs to stop word list. Then other words were identified as tokens which would be sent to dictionary for translation. Some of the resulting tokens were not translated by dictionary, because they were the inflected form of words or they were plurals. In this research, morphological analysis, which is a post-translation technique, was done on those words which were not translated by the dictionary. By using morphological analysis technique, those words changed to normalized forms and we tried to look them up again in the dictionary. Then translated words were added to the final English queries.

We also examined phrasal translation and compared it with the word-by-word translation. Before phrasal translation, we needed to identify probable phrases in queries. So part-of-speech tagging was applied on query terms. In this process, each word in a given query was assigned a tag which showed its grammatical class. Then by observing Persian language phrase structure patterns, potential phrases were extracted and translated as phrases.

Resulting queries were then made available to some searchers and they were asked to retrieve relevant documents from the Google search engine. In retrieval systems like Google that have a very large database, each query returns huge number of documents, sometimes millions of documents. It is clear that the evaluation of relevance for all documents is impossible, so a sampling method called 'pooling' introduced by TREC[2] was used. In this method, a pool with the depth of 100 records is made for each query. The pools are made by listing retrieved documents which are at the top of returned lists. We judged relevance of documents in the pools by using relevance criteria, proposed in the narration field of original queries. Relevance judgments were carried out by using a binary method in which we assigned 1 for *relevant* documents and 0 for *irrelevant* ones. By using these scores, Mean Average Precision (MAP) and precision at different cut-off levels for each retrieval approach was measured. The higher the MAP score, the more effectiveness of retrieval approach. Voorhees (2003) says" this method of CLIR evaluation has shown its efficiency in several experiments".

Results

To study the effect of applying *NLP* techniques on the efficiency of Persian-English *CLIR*, we used a dictionary approach and evaluated the degree of *NLP* processing impact

on the *CLIR* system performance. First, the impact of morphological analysis on *CLIR* effectiveness was measured. This showed that morphological analysis of words which were not translated by the dictionary increased the MAP scores.

Table 1
Retrieval Effectiveness of CLIR with and without Morphological Analysis

CLIR approach	Mean average precision	
CLIR without morphological analysis	0/180	-
CLIR with morphological analysis	0/223	%23

In Table 1, the results of measuring MAP scores for queries translated with and without morphological analysis are summarized. The MAP score for morphologically analyzed queries is better than those without that analysis. It yields %23 more effectiveness. The same results were obtained when we measured retrieval effectiveness at different cut-off levels.

Table 2
Retrieval Effectiveness of CLIR with and without Morphological Analysis at Different Cut- off Levels

precision	*CLIR* with morphological analysis	*CLIR* without morphological analysis
At 5 docs	0/485	0/396
At 10 docs	0/441	0/372
At 20 docs	0/399	0/310
At 30 docs	0/318	0/263

Overall results show that morphological analysis of the words not translated by the dictionary can improve *CLIR* effectiveness. Mapping inflected words or plurals to normalized forms may produce translations which would increase the MAP score of resulting queries.

Other findings in this research showed that phrase translation, in comparison to word-by-word translation, resulted in more efficiency in Persian-English *CLIR*.

Table 3
Retrieval Effectiveness of CLIR with Phrasal and Word-by-Word Translation

CLIR approach	Mean average precision	
CLIR with word by word translation	0/223	-
CLIR with phrasal translation	0/319	%43

Results shown in the above table reveal that the map scores of these two translation

methods were 0.319 for phrasal and 0.223 for word–by-word translation. It indicates that identifying phrases in Persian queries and translating them as a semantic unit improves *CLIR* effectiveness by %43(compare this with using word-by-word translation method).

Table 4

Retrieval Effectiveness of CLIR with Phrasal and Word-by-Word Translation at Different Cut-off Levels

precision	Phrasal translation	Word-by-word translation
At 5 docs	0/592	0/485
At 10 docs	0/541	0/441
At 20 docs	0/512	0/399
At 30 docs	0/481	0/318

The above results show that Precision for phrasal translation of queries at different levels of retrieval is clearly higher than the word-by-word translation. This finding combined with the previous results, justify the use of part-of-speech tagging technique in detecting phrases in Persian queries which could in turn be used in a phrasal translation method.

Discussion

Inflected form of words is used for expressing grammatical information about time, quantity and gender. Pirkola (2001) made mention of inflected forms as issues in *CLIR* translation which must be resolved. Morphological analysis used in this research showed its efficiency in resolving the problems arising from the words not translated. The number of research query terms used in this research showed that from a total of 266 query terms, 128 words were not in the dictionary. Exclusion of this large number of words would mean that in the final queries about %48 of source query terms are not included.

Besides, most of queries are made up of phrases whose translation is a difficult task. This is a serious problem for CLIR especially in languages like Persian which use phrases to communicate meanings and ideas. The findings of this research show that using *NLP* technique of part-of-speech tagging is a good way to identify phrases in queries. By the use of a translation resource which has a good coverage of phrases, *CLIR* effectiveness will increase and the number of irrelevant retrieved documents will decrease. Our findings are in congruence with those of other researchers who also maintained that phrasal translation was an appropriate method for query translation.

Conclusion

There are many reasons why a *CLIR* system does not result in a good retrieval performance as a monolingual IR does. The first and most important reason is the existence

of two different languages in *CLIR* which have the*ir* separate structures and vocabularies. This dichotomy causes ambiguity in translating *CLIR* queries. It is a problem that monolingual information retrieval would never encounter.

NLP techniques such as tokenization, part-of-speech tagging and the use of morphological analysis can address the CLIR translation problem. Phrasal translation along with query terms not translated are even more problematic. Multi-term concepts called phrases are easily translated via MRD when it is empowered by NLP tools. In this study, we have shown that NLP techniques are useful aids for the purposes of CLIR. The creation of such tools in Persian therefore is a task that calls for action on the part of researchers in this field.

Endnotes

1. Available at: www.farsidic.com
2. Text Retrieval Conference

References

Adriani, M. (2000). Using statistical term similarity for sense disambiguation in cross-language information retrieval. *Information retrieval, 38* (2), 69-80.

Alizadeh, H. (2004). Problems of information access in the world of networks. Faslnameh Ketab, *15* (2), 115-121.

Aljlayl, M.& Phir, F. (2001). *Effective Arabic- English Cross- Language Information Retrieval via_Machine- Readable Dictionaries and Machine Translation*. Oral presented at the ACM Tenth Conference on Information and Knowledge Management, Atlanta.

Ballesteros, L. & Croft, B. (1998). Resolving Ambiguity for Cross- Language Retrieval. *SIGIR,* 64-71.

Chen, H. H. (2002). Chinese information extraction techniques. SSIMP- 2002.

Davarpanah, M. R., Sanji, M. & Aramideh, M. (2009). Farsi lexical analysis and stop word list. *Library Hi, 27* (3), 435-449.

Edwards, J. (1994). M*ultilingualism*. London: Penguin

Hedlund, T. (2003). An Extendable Query Translation System. Paper Presented at the ACM SIGIR Workshop for Cross language Information Retrieval, 2003.

Hull, D. & Grefenstette, G. (1996). Querying Across Languages: A Dictionary –Based Approach to Multilingual Information Retrieval. In *Proceedings of the 19th Annual International ACM Sigir.* Zurich, Switzerland, 49-57.

Lyons, J. (1981). *Language and linguistics: An introduction.* Cambridge: Cambridge University press.

Mehrad, J. & Naseri, M. (2008). *Natural language processing and information retrieval.*

Tehran: Chapar.

Pirkola, A. (2001). Dictionary – based cross- language information retrieval: Problems, methods and research findings. *Information Retrieval, 4 (*4C3), 209-230.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14,* 130-137.

Sheridan, P. & Smeaton, A. F. (1992).The Application of Morph-Syntactic language processing to effective phrase matching. *Information processing and Management*, *28*(3).

Voorhees, E. (2003). Overview of TREC2002. Retrieved January 21, 2006, from http://-nlpir.nist.gov.