Query Expansion for Cross Language Information Retrieval Improvement

Benoît Gaillard, Jean-Léon Bouraoui, Emilie Guimier de Neef, Malek Boualem
Orange Labs
Lannion, France
{benoit.gaillard, jeanleon.bouraoui, emilie.guimierdeneef, malek.boualem}@orange-ftgroup.com

Abstract— **This paper is devoted to a new method that uses query expansion to improve multilingual information retrieval. The backbone is an Information Retrieval (IR) system based on a search engine and a multilingual module based on statistical machine translation of documents. To this system is added a Query Expansion (QE) module which mainly uses linguistic resources to perform the expansion. The aim is to use QE to overcome the limitations of machine translation, and to retrieve more relevant results. The authors demonstrate, with examples, the usefulness of such a system. They also validate it with several measures, which show a clear reduction of the silence for results.**

Keywords-component; Cross-Language Information Retrieval, Machine Translation, Query Expansion, Natural Language Processing

I. Introduction

Nowadays, the number of documents potentially accessible is constantly growing, in more and more different languages. However standard queries, containing 2-3 terms in average, are less and less likely to be sufficient to retrieve all of the relevant documents.

Consequently, advanced techniques are necessary to enhance the performances of Information Retrieval (IR) systems, such as Cross Language Information Retrieval (CLIR). It enables searching on multilingual sets of documents, involving languages that might be unknown to the user. Bridging the gap between document and query languages requires the application of machine translation techniques to queries, indices, or both. Translation software can induce linguistic differences between translated data and human language. Our proposal is to overcome this problem by using Query Expansion (QE). QE consists in adding new words to the initial query. Thus, the query matches documents that do not contain terms from the initial query. In a nutshell, our key idea is to combine QE with CLIR in a Multilingual Multimedia Information Retrieval (MMIR) prototype.

The paper presents an implementation of this idea. In a first part, we describe the main principles and applications of CLIR and of QE. We also present how QE can address the issues expressed above, and we demonstrate the interest of the method we propose. Then, we describe how we implemented this solution. Finally, we validate the system and comment the obtained results.

II. CLIR approaches for multimedia Information Retrieval

In this section, we first introduce Cross Language Information Retrieval (CLIR). Then we describe the multimedia context of the MMIR prototype, on which this paper is based. We justify the specific approach chosen for the CLIR application and finally, we point at remaining issues that are inherent to this approach.

A. Cross Language Information Retrieval

CLIR search engines enable users to retrieve content in a language different from the language used to formulate the query. Since the variety of online content languages is increasing, internet users need tools to gain access to this multilingual information. CLIR attempts at fulfilling this need. To reach this aim, it provides search engines that enable users to enter a query in their own language, in order to retrieve lists of results in a different language. CLIR systems are often based on machine translation techniques, combined with a regular monolingual search engine. They fall into two main categories: translating queries, on one hand, or translating indexed data, on the other hand.

B. The MMIR prototype

The first version of the MMIR prototype involves short videos in the domain of news. The videos are selected from online web TV channels, from UGC portals, or from online news agencies. The first languages considered by the application are French and English. More precisely, the queries are supposed to be written in French and the documents are originally in English or French.

1) Indexing metadata: Indexed data usually consists in information that is associated to videos in order to describe them, and is referred to as "metadata". For example, titles, subtitles and scenarios of movies are all metadata. If the video is broadcasted by a TV channel, significant amounts of additional metadata, such as the EPG (Electronic Program Guide) are available for indexing. In the case of the MMIR prototype, the indexed metadata consists in two fields: the title of the video and a short description, of approximately 5 sentences.

C. *Translation uses in Cross Language Multimedia Information Retrieval*

1) Translation of indexed content vs. translation of queries: in contrast with queries, the content of indexes lends itself naturally to machine translation, because it consists in grammatical textual data. Moreover, its theme is often well defined, which is helpful for machine translation because it enables the building of parallel bilingual corpora for training. That is why comparisons of content translation approaches with query translation ones ([1], [2]) tend to favor content translation.

2) Query translation as a pragmatic choice: nevertheless, most multilingual search engines are based on query translation. The well known Google search engine, for example, simply translates queries and then perform monolingual search using the translated query. This is due to the dramatic growth of the quantity of data to be indexed. Translating all this data into the significant number of languages to be taken into account would require amounts of storage and computing power that are beyond the means of most applications. Conversely, merely translating the queries and the documents that are chosen by users is much more cost effective. As a consequence, this approach has been focused on from the beginning of CLIR research [3] to more recent investigations such as [4].

D. *Metada translation for the MMIR prototype*

In the MMIR prototype, the nature and theme of information is well defined, the quantity of metadata is relatively manageable, and only one language pair is to be considered. As a consequence, translating the metadata-based content before the indexation is possible. We have shown above that the approach to CLIR by translating documents before indexation is, when possible, potentially more accurate than the approach involving query translation. As a consequence, in the context of the MMIR prototype, CLIR has been based on the translation of documents, despite the fact that most CLIR systems are based on query translation. This paper proposes a method for optimizing the use of translated documents in this context

E. *Pitfalls of the document translation approach*

1) Missing words in the translated text. The structures and vocabulary of texts that were generated by machine translation software are different from those of human generated texts. Such differences can arise from the limits of statistical machine translation. Infrequent structures can hardly be learned by statistical methods, and exceptional features of language will therefore not appear in translated text. For example, Google translates[1] “the apple eats the child” by “l’enfant mange la pomme” which means “the child eats the apple”. Another cause of discrepancy between MT generated text and human generated text is the management of words bearing multiple meanings. In the context of the MMIR prototype, the issue can be illustrated with some examples. The English word “lock” can be translated into (at least) three French words of complete different meanings: “écluse” (meaning a dam on a canal that enables boats to go up and down different levels), “cadenas/serrure” (meaning something to fast something, a padlock), or “enfermer” (meaning to secure someone or something in something or somewhere). This word can also mean a specific hair style. In the MMIR prototype, tested in august 2009, the most frequent translation of “lock” is “écluse” and as a consequence, in some translated video descriptions or titles Jennifer Anniston has dams in her hair, some bicycles are secured by dams… Beside these unavoidable but anecdotic mistakes, a more serious consequence is that a user searching with the query “antivol” (meaning padlock) or “mèche” (tuft of hair) will not find the aforementioned relevant results in the retrieved list.

The same issue applies to synonyms. For example, “night club” can be translated into French by “boîte de nuit” or by “discothèque”, which share the same meaning. The translation software solution used for MMIR prototype systematically translates it by “boîte de nuit”, which is correct. As a consequence, users can not find any relevant result upon querying with “discothèque”, which also is a frequently used word in the French language to denote night clubs. This issue is illustrated in Fig 1.

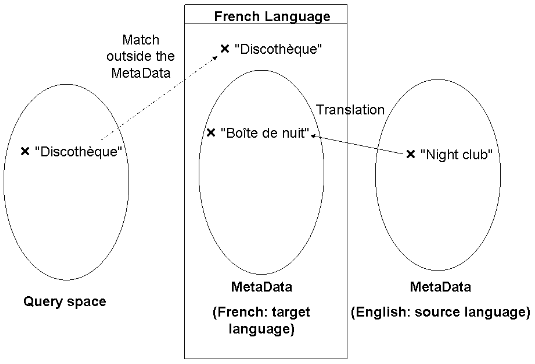

Figure 1. Miss-match between queries and translated metadata

2) Quantitative evidence: these examples show that the vocabulary of translated documents does not contain the general vocabulary of the query language. In order to measure how valid this assertion is, we evaluated the number of words in a corpus of queries that do not belong to the corpus of translated metadata (titles and description of the indexed videos).

The corpus of queries consists in 51461 different queries (corresponding to 50973 different words[2]) submitted to OPF (Orange Portal France) for searches on the theme of news, in January 2009. The corpus of metadata consists in titles and

[1] Google (cf. http://www.google.fr/language_tools) uses statistical machine translation.

[2] This count includes non-lexical terms such as URLs, mails, etc.

short descriptions (approximately 5 sentences) of about 57000 videos coming from a diversity of online broadcast channels. It was translated by a statistical machine translation service trained on the Europarl bilingual corpus. Statistical machine translation techniques are introduced for example in [5].

A significant proportion of queries is composed of words with spelling mistakes, and would obviously not belong to corpora of translated documents. Processing them in order to match the index contents is beyond the scope of this article; therefore, they are not taken into account in this evaluation. For similar reasons, URLs and named entities are not taken into account either. Our aim is to only consider queries that consist in correctly spelt French words. Thus, the query corpus was intersected with two French lexicons: Lexique 3[3], made of 135 000 words and a lexicon extracted from the Orange thesaurus for Tilt [6].

This selection process resulted in a set of queries that we call "usable corpus" of $N_{usable} = 34070$ different French words,. Out of this "usable corpus", $N_{exo} = 2800$ words, such as "mincir", "internaute", "alouette", "mairesse" or "potager" do not belong to the corpus of translated metadata, so we have a "mismatch ratio" of:

$$R_{missmatch} = \frac{N_{exo}}{N_{usable}} = \frac{2810}{34070} = 8\%$$

This demonstrates that adjusting the query space to the translated metadata space would be beneficial to the CLIR prototype. The next section is devoted to the presentation of our solution to perform this adjustment.

III. Adjusting the Query Space to the Data Space using Query Expansion

A. *Query expansion for CLIR*

1) Introduction to query expansion. Query expansion (QE) is an Information Retrieval (IR) technique. Its aim is to improve the relevance and quantity of the results retrieved by IR systems. It starts with the observation that many queries do not return the whole set of relevant documents. Indeed, there usually is an inconsistency, stated and measured by several works (see for example [7] or [8]), between the queries and the corresponding indexed documents. QE overcomes this issue; it consists in adding to the original query, new terms related to it, or even to reformulate it. For instance, let "house" be the initial query Q1. If QE technique is applied to Q1, it would become Q2, containing "house OR lodge OR domicile (…)".

The QE process includes two major phases. The first phase consists in searching "related" terms to add to the query. These terms will be candidates for the expansion. The second phase, the expansion itself consists in integrating the expanded terms into a new query. We outline the first phase below.

There are many different methods for choosing terms to include in the expansion. To summarize, they differ according to the resources they are based on:

- Results from a previous search: the user input an initial query, and indicates which of the retrieved documents are the most relevant. From this set, some keywords are extracted, and used to carry out a second search. Automated versions of this method are named "pseudo-relevance feedback" or "blind relevance feedback".
- "Knowledge structures": we mean by this term any resource which is independent from the results obtained with the initial request. These resources can consist on the index of the considered IR system, or at least a part of it. The resources can also be external to the documents indexed by the IR system: lexical and/or semantics thesaurus, ontologies, etc. Here, the candidate terms for the expansion are selected on the basis of semantic, lexical, and/or morphological relatedness. The literature shows that potentially any type of document can serve as external resource.

QE is overall recognized as a useful technique, which has been used for more than three decades. Its main flaw is the risk of "query drift". This term refers to a deviation from the original intent of the user. It occurs when the original meaning of the initial query is distorted by the expansion. For example (from [7] (p. 45)), let us consider a user searching documents about the term "Nirvana", in its Buddhist acceptation. A bad application of QE could transform the query into "nirvana kurt cobain live band music". This would be a typical case of "query drift", because the meaning of the new query, as well as the results it would retrieve, would have no relation to the original intent of the user. It entails that none of the retrieved documents is relevant to the initial search.

Query Expansion as a bridge from non-matching queries to translated documents space. From the previous description, it appears that QE shares many common points with CLIR. Notably, both these techniques aim at retrieving more relevant results to a given query. Also, they both consist in processing the queries or the indices to add more information. Therefore, combining these techniques seems promising, as confirmed by some previous studies carried out on this topic ([9], [10]). These works mostly use QE to disambiguate results obtained with CLIR or to lower the impact of translation errors ([11]). In our work, our aim is to use QE in order to address the problems spotted in section II.E above.

Our assumption is that QE will take into account some nuances of the terms, which are otherwise ignored by the single-word translation. For example, let us consider the French word "discothèque" used as an example in section II.E.1). Ideally, QE would give for this word terms such as "boîte de nuit". Consequently, it would allow returning documents containing other synonyms for "discothèque".

The principle consists in expanding each query, so that more query words belong to the vocabulary of the translated documents. The simplest version of the method applies plain query expansion to each query, before performing the cross

[3] Available on www.lexique.org.

language search. This principle is illustrated by fig. 2 below: Each cross corresponds to a word. The “A” cross corresponds to the initial query, and the “A’”, “A’’”, etc. crosses to its expansions. One of the expansion matches with one word from the metadata, whereas it wasn’t the case with the initial query.

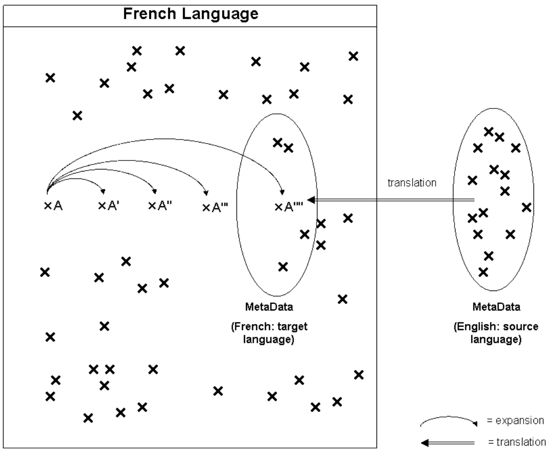

Figure 2. QE as solution to the CLIR miss-match problem

B. *Merging QE and CLIR into a single prototype*

1) CLIR module. The CLIR module consists in adding a translation service to the IR engine. The prototype on which this paper is based involves translating the contents before indexing them. Only the translation is indexed, whereas the original text is kept in the memory of the engine, in order to present the original version of texts to users.

2) QE module. It is based on the TiLT platform[4], designed at Orange Labs. Its aim is to propose, for one or several query terms, some corresponding expansion terms. For now[5] , it can perform QE according to five modes:

- Inflection: case, gender, number, tense, person, mood, or voice;
- Synonyms: the synonyms of the terms from the initial queries;
- Hyperonyms: in the sense of usual linguistic definition: (general to specific);
- Derivated terms: any term that is semantically related to the initial query;
- Geographical expansion: this mode allows obtaining, for a given name of an area, the name of the areas that are included in it.

The expansion terms are produced, according to these modes, through the use of a thesaurus, built in Orange Labs. It is composed of about 100 000 items, each one associated to several linguistic features (part of speech, sense, etc.). We also used the Geonames database [6] to build the geographical expansion mode.

3) IR system. The system is based on the Orange Labs search engine. It performs the tasks of indexing a collection of documents, and of searching through it. It accepts simple but also complex queries (i.e. made of several terms linked to each other by complex Boolean operations (AND, OR, etc.)).

The backbone of the architecture is the indexing and search engine, which makes the connection between the CLIR and QE modules. CLIR uses the engine to search the collection of documents. The QE module expands the initial query, and then the search engine searches the index of translated metadata with the expanded query. These three components (CLIR and QE modules, search engine) are combined into the architecture displayed in fig. 3.

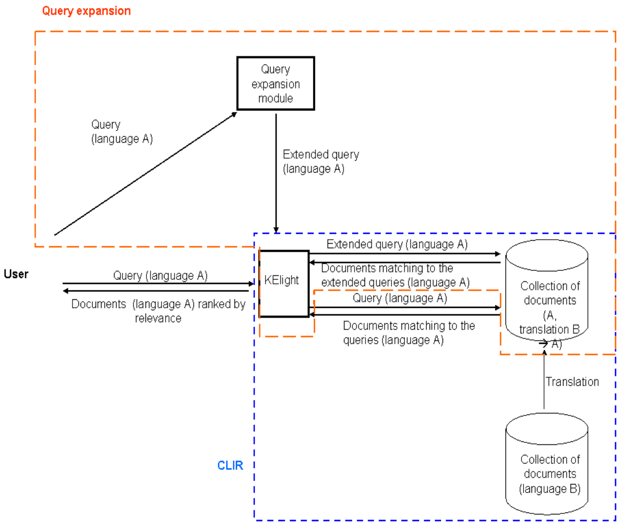
Query expansion
Query expansion module
Query (language A)
Extended query (language A)
User
Query (language A)
Documents (language A) ranked by relevance
KElight
Extended query (language A)
Documents matching to the extended queries (language A)
Query (language A)
Documents matching to the queries (language A)
Collection of documents (A, translation B → A)
Translation
CLIR
Collection of documents (language B)

Figure 3. Synoptic Architecture of the QE-CLIR prototype

IV. Method Validation

In this section we use the OPF set of queries described in section II.E.2. This corpus is used in two different ways. In section 4.A we use the whole unfiltered corpus, in order to evaluate the scope of QE. In section 4.B however we filtered the corpus, selecting only single words queries, in order to provide more accurate measures of silence reduction.

A. *Range of query expansion*

Some queries are not suitable for expansion whereas some others get expanded into a significant number of alternatives. The following measures evaluate how much expansion is performed on the corpus of queries:

[4] *TiLT* is a “multilingual Natural Language Processing platform” (cf. [6]).

[55] We plan to add other modes soon, notably to expand from and to Named Entities (proper nouns …). Only geographical Named Entities are carried out for now.

[6] www.geonames.org. This is an open source (under Creative Commons licence) database of worldwide geographical information. Provided details are numerous: names of the area in different languages, number of inhabitants, latitude and longitude, towns, etc.

- The number of expanded queries divided by the number of queries: $ER =$ 0.59. The remaining 40% of queries that are not expanded mainly consist in named entities, URLs or orthographic mistakes.

- The mean number of words in the expansion: $\langle N_e \rangle = 29$. This is a significant number, considering it takes into account the 40% zero word expansions.

- The mean number of words in the expansion, provided the query is expanded: $\langle N_{e+} \rangle =$ 48. The average number of words in expansions is about 50 words but this figure varies a lot, from a couple of words to a few hundreds.

- The mean number of words in the expansion, provided the query is expanded, and provided the query is not of geographical nature: $\langle N_{e+,notgéo} \rangle =$ 34.67. We distinguish the geographical expansions, since they usually lead to many more results than the other modes of expansion.

B. Silence reduction

The first aim of QE for search in translated documents is to solve the issue related to a significant level of silence, as exposed in section III.A.1. This section shows the success of QE at reducing silence. Firstly, we consider initial queries that do not enable the search engine to retrieve any result at all. From theses queries, we show that QE enables the search engine to retrieve several relevant results.

The search is performed, with the MMIR prototype, on the basis of a Boolean "OR" query, the clauses of which are the alternatives proposed by the expansion. Queries are expanded into a number of alternatives that varies between zero and several dozens. The expanded query consists in a combination of the alternatives, separated from each other by the Boolean operator "OR". For instance, the initial query "house" becomes the expanded query: "house OR lodge OR domicile (…)".

1) Illustration of the benefits of QE for CLIR. Here we list queries that illustrate the benefits of the method. The examples given in Table 1 are in French, since it is the target language of the application.

Table 1. Samples of expansions that allow to access to results

Initial Query	**Expanded Query**	**Number of Results**
mincir	amincir,amincissement,mince, minceur,minci,mincie,mincies, mincis,mincissant,svelte,sveltesse	35
auvergne	allier,auvergnat,cantal,haute-loire,puy-de-dôme	1
ciné	cine, cinoche,ciné,cinéma,cinés,salle-de-cinéma,salle-obscure	109
mômes	bambin,enfance,enfant,enfantin,gamin,gosse (…)	1491

2) Quantitative Evaluation of Silence reduction. Two measures evaluate silence reduction. For technical reasons, both of them were computed on 1300 queries that contain a single word, extracted from the usable OPF corpus that was described in section II.E.2.

- Number of queries for which the search engine does not find any result, whereas results are found for their expansion: $N_{s=0}^{\exp>0} = 121$. This number has to be put into perspective because it is computed on the 1300 single word queries extracted from the "usable corpus": about 10% of queries are taken into account. Besides, only 237 queries from these 1300 do not produce any results without expansion. If we consider this subset of queries, it means that the expansion has positive effect for more than 50% of them!

- Average number of results added by expansion: $\langle N_{\exp} - N_s \rangle = 295.79$. This number is larger than we expected initially. It is explained by the fact that some of the expansions, though relevant, correspond to widely used words; thus, the number of documents that contain them is high. Let us consider the query "parole" ("speech" in English); it gives 22 different expanded words. Among these expansions, is the word "dit" ("says" or "said" in English): the expansion is relevant, but gives numerous results. Indeed, many documents are related to the fact that someone said something.

C. Interpretation of results and comments

More than half of the queries of the complete OPF corpus, including words with spelling mistakes, URLs, etc. are expanded into several choices. This means that most of the queries of the "usable corpus" are expanded. The average number of words proposed by the expansion is larger than the maximum number of alternatives that the Orange search engine could handle in a Boolean query. It is worth noticing that the quantity of words provided by QE is unevenly spread across different queries. Some of them are only expanded into a couple words, whereas others are into hundreds of words. This shows that optimizing QE for CLIR does not depend on increasing the quantity of proposed words, but more on the fine adjusting of the process to specific words or contexts.

Silence reduction, which is the first aim of the proposed method, is significant as the results in IV.B show. Furthermore, our first overview of the QE results does not show much query drift. Therefore the precision is unlikely to be drastically affected. Indeed, the expansions shown in section IV.B.1 do maintain the meaning of the original query.

One can observe very large discrepancies in silence reduction according to which query is considered. For example, the expansion of the query "Auvergne" only retrieves 1 answer, whereas the expansion of the query "mômes" retrieves 1500 answers. This can not be explained by query drift or loss of precision, because the expansion of "môme" maintains its general sense. Similarly, the impressive silence reduction achieved with the query "parole" can be explained by its expansion into the widespread word “dit”. The small number of retrieved documents for “Auvergne” can be explained by the fact that very few videos about Auvergne are broadcasted by English or American channel. However, this explanation does not hold for queries like “potager”, for which the expansion does not retrieve any result. The reason for this discrepancy must be that, for some queries, even QE fails to reach the translated metadata space. For example, the English phrase that should be translated into “potager” is "vegetable garden" but unfortunately it is translated by “jardin de légumes”, which is not proposed by expansion.

V. Conclusion and future developments

This article presented a novel approach for the CLIR technique that consists in indexing translated documents. This technique can provide better results than query translation. However, it can induce some difficulties, such as the fact that the translation of a word can systematically ignore some of its relevant senses or alternative translations. Our key idea is to use Query Expansion (QE) to overcome this problem, by adding similar senses to the initial query.

We presented the solution both from a theoretical and practical perspective. Our first results highlight a success at reducing silence through many examples extracted from a corpus of queries coming from real, public usage. As a first step we limited ourselves to single word queries but the improvement is already clear.

Further developments will be added to this work, according to two different directions. First, more complete measures will be added to our evaluation methodology. We will compare the precision of the CLIR system with or without QE. We will also extend the evaluations presented in the current paper to queries containing several words. The second line of developments consists in designing more adapted QE processes. In the current paper the QE is performed independently from the target corpus, i.e. the corpus of translated documents. In subsequent versions of the method we will design solutions to tune the QE process so that the selected words for the expansion will be more precisely adapted to the given application.

Let’s conclude on a more global consideration: our work shows that, in order to improve efficiency of a search engine, the solution seems to involve a variety of techniques to leverage as much context as possible in order to optimize possible interpretations of the few words provided by queries.

References

[1] Oard, D. : "A comparative study of query and document translation for cross-language information retrieval" in "Proc. of the 3d conference of the Association for Machine Translation of the Americas on machine translation and the information soup", Farwell, D., Gerber, L. and Hovy, E. (Eds), Springer –Verlag, pp. 472- 483, 1998.

[2] Clough, P.: "Caption and Query translation for Cross-Language Image Retrieval", in "Proceedings of CLEF 2004", Peters C. & al. (Publishers), CLEF 2004, pp. 614—625, Springer-Verlag Berlin Heidelberg,(2005.

[3] Yang, Y., Carbonell, J. G., Brown, R. D., Frederking, R. E.: “Translingual Information Retrieval: Learning from bilingual corpora”, Artificial Intelligence, 103, pp 323—345, 1998.

[4] Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, R. B., Hiemstra, D. and DE Jong, F. M. G.: "WikiTranslate: "Query Translation for Cross-lingual Information Retrieval using only Wikipedia", in working notes CLEF, 2008.

[5] Och, F. J., Ney, H.: “Statistical Machine Translation”. EAMT Workshop, pp. 39-46, Ljubljana, Slovenia, May 2000.

[6] Heinecke J., Smits G., Chardenon C., Guimier De Neef E., Maillebuau E., Boualem M.: « TiLT: plate-forme pour le traitement automatique des langues naturelles », Traitement Automatique des Langues 2008 Volume 49 Numéro 2, pp. 17-41, 2008.

[7] Billerbeck B.: “Efficient Query Expansion”, PhD Thesis, RMIT University, Melbourne, Australia, September 2005.

[8] Xu J., Croft W. B.: “Improving the Effectiveness of Information Retrieval with Local Context Analysis”. ACM Transactions on Information Systems, Vol. 18, No. 1, pp. 79-112, January 2000.

[9] Bellaachia A., Amor-Tijani G.: “Enhanced Query Expansion in English-Arabic CLIR,” pp. 61-66, 19th International Conference on Database and Expert Systems Application, 2008.

[10] Darwish K., Oard D.: “CLIR Experiments at Maryland for TREC2002: Evidence combination for Arabic-English retrieval”. Proceedings of the Text Retrieval and Evaluation Conference (TREC 2003), 200

[11] Ballesteros L., Croft W. B.: “Resolving ambiguity for cross language retrieval”. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 6471, 1998.