Structured Indexing Model for Cross-Language Information Retrieval

Chedi Bechikh Ali* — **Hatem Haddad****

* *ISG, Université de Tunis*
LISI laboratory, INSAT, Carthage University
** *Mevlana University, Konya, Turkey*
$\{chedi.bechikh, haddad.hatem\}$ @$gmail.com$

ABSTRACT. *In recent digital library systems or World Wide Web environment, parallel corpora are used by many applications (Natural Language Processing, machine translation, terminology extraction, etc.). This paper presents a new cross-language information retrieval model based on the language modeling. The model avoids query and/or document translation or the use of external resources. It proposes a structured indexing schema of multilingual documents by combining a keywords model and a keyphrases model. Applied on parallel collections, a query, in one language, can retrieve documents in the same language as well as documents on other languages. Promising results are reported on the MuchMore parallel collection (German language and English language).*

RÉSUMÉ. *Dans les systèmes récents de bibliothèques numériques ou dans le contexte du Web, les corpus parallèles sont utilisés par de nombreuses applications (traitement du langage naturel, la traduction automatique, extraction de terminologie, etc.). Cet article présente un nouveau modèle de recherche d'information inter-langue basé sur le modèle de langue. Le modèle évite la traduction des requêtes et/ou des documents ainsi que l'utilisation des ressources externes. Il propose un schéma d'indexation structurée des documents multilingues en combinant un modèle de mots-clés et un modèle de phrase-clés. Appliquée sur une collection parallèle, une requête dans une langue, peut récupérer des documents dans la même langue ainsi que des documents dans d'autres langues. Appliqué à la collection parallèle MuchMore (en langue allemande et en langue anglaise), le modèle a montré des résultats prometteurs.*

KEYWORDS: *Cross language information retrieval, parallel collection, multilingual index, keyphrases, natural language processing.*

MOTS-CLÉS : *Recherche d'information inter-langue, collection parallèle, index multilingue, phrase-clés, traitement de la langue naturelle.*

1. Introduction

Cross-language information retrieval (CLIR) allows an information seeker to apply a request in one language and to find information in a different language. Accordingly, multilingual information research models must overcome and eliminate language barriers by allowing an Information Retrieval System (IRS) to retrieve relevant documents expressed in languages other than the query language. For example, considering the German query [1]: *"Arthroskopische Behandlung bei Kreuzbandverletzungen"* [2], the following English text [3] should be judged relevant by the CLIR system: *"... In all knee ligament procedures, arthroscopy is obligatory for diagnosting and conducting meniscus surgery ..."*.

Most of the techniques proposed to solve the problem of Cross-Language Retrieval center around a common idea: they attempt to translate the query from the user's language to the language of the documents (Peters *et al.*, 2012). In most cases, the translation is done using a dictionary or a machine translation system. Any given word may have multiple possible translations, so significant effort has been devoted to disambiguating the resulting translations.

CLIR is facing the correct terminology equivalent (translation) selection challenge from one language to another. Indeed, the terms of the various languages almost never cover the same semantic field and the sense drift is unavoidable in a translation. To minimze the sense drift and the effort on the translation we propose to avoid the query or the document translation step.

Further, the CLIR systems are facing the challenge to be effective and to rank the relevant retrieved documents. To achieve this, CLIR systems must use an expressive documents (as well as queries) representation. Earlier works showed that the use of simple words as keywords is not always accurate enough to represent the documents contents due to the words ambiguity (Arampatzis *et al.*, 1998). Thus, we propose a documents representation and queries representation using keyphrases. By keyphrase, we refer to a list of phrases of two or more words such as many academic journals include in the beginning of an article (Bechikh-Ali *et al.*, 2015). Keyphrases can be extracted statistically, linguistically, or by combining the two approaches (Haddad and Bechikh-Ali, 2014). We note that the keyphrases extracted statistically to represent the document content may contain noise that may affect negatively the IRS performances. Considering the following example:*"In all knee ligament procedures, arthroscopy is obligatory for diagnosting and conducting meniscus surgery ..."*. The statistical methods extract keyphrases such as "in all", "is obligatory", " for diagnosting", etc. These keyphrases can decrease the IRS performance. On the other hand, linguistic methods extract keyphrases such as: "knee ligament", "meniscus surgery", etc. We note that these keyphrases are more likely to represent the content or the topic of the sen-

1. German Query number 1 from the MuchMore collection (**?**)
2. Arthroscopic treatment of cruciate ligament injuries
3. English document DerChirurg/70681093 from the MuchMore collection

tence than those extracted by statistical methods (Bechikh-Ali *et al.*, 2015). For this reason, in addition to keywords, we propose to represent documents and queries by keyphrases.

The remainder of this paper is organized as follows. The related works are reviewed and discussed in Section 2. In Section 3, we introduce our proposed CLIR model based on a structured multilingual indexing. The experimental setup is detailed in Section 4. Evaluation results and discussion are given in Section 5 followed by a conclusion and future improvements to end the paper.

2. Related Work

Four types of approaches for CLIR are identified (Oard and Diekema, 1998) and can be divided into two categories (Kishida, 2005):

– Translation: in this category the following approaches can be used: query translation (QT) (translate the query representation to match the document representations), document translation (DT) (translate the document representations to match the query representation) and interlingual techniques (IT) (translate the document and the query representations into a third language or semantic space). IT approach uses a language independent representation for both queries and documents of a given parallel document collection (Chew and Abdelali, 2007). QT and DT use dictionary-based translation (i.e. machine-readable, ontologies, dictionaries) or machine translation (Zhou *et al.*, 2012). In this context, CLIR is an integration of words translation into word-based retrieval models (Hieber and Riezler, 2015).

– No translation: in this category, the cognate matching is used between languages having a close linguistic relationship (English, French, Spanish, ...) where unchanged words (such as proper nouns or technical terminology) can be expected to match successfully (Oard and Diekema, 1998; Gey, 2005).

QT has become the popular matching approach for CLIR due to the fact that it is less computationally costly to process the translation of a query than the translation of a large documents set. On the other hand, it presents two main disadvantages. The first disadvantage is the word translation ambiguity. The reason is that queries are often short and provide little context for disambiguation leading for many translation possibilities (Oard and Diekema, 1998). For example, "air flight'" will be translated to French as "vol air". The word “vol” can refer to either "roberry” or to "flight". Furthermore, the dictionary-based translation approaches are limited by the availability of phrases and compound words for translation.

Various methods have been proposed to solve the problems of disambiguation, often relying on the document collection or integrating a translation step directly into the research model (Kraaij *et al.*, 2003). Other methods rely on external resources, such as Wikipedia (Nguyen *et al.*, 2008) or the Web (Hu *et al.*, 2008). Gao et al. in (Gao *et al.*, 2006) proposed syntactic translation models to address the homonymy problems. Translations candidates proposed by these models are then reclassified us-

ing the learned model to minimize translation error. Because of modeling translation and modeling retrieval separately, (Ture *et al.*, 2011) proposed to include a weighted translation into the query structure. Hiemstra et al. (Hiemstra and de Jong, 1999a) applied QT using one translation per source language query word. QT using all possible translations per source language query word. They concluded that using all possible translations for searching leads to better average precision retrieval performance than using one translation. In the same way, (Nguyen *et al.*, 2008) mapped queries to Wikipedia concepts and used corresponding translations of these concepts in the target language to create the final query.

The second disadvantage of the QT approach is, considering a N-lingual environment, the number of required systems to achieve CLIR reaches $\frac{N(N-1)}{2}$ (Costa-jussà M, 2014). For this reason, most research in CLIR has not attempted to use many languages at a time. The reason is, that for each new language pair, a new translation algorithm must be used.

When using QT approach (and also in the case for DT strategy), once queries (and also the case for documents) are translated, a monolingual information retrieval system is then used. The quality of the IR step is strongly correlated to the quality of the translation step (Hieber and Riezler, 2015).

Other researchers have attempted to apply the language model (LM) to CLIR tasks. One of the advantages of the LM approach to CLIR tasks is to enable researchers to put translation probability linked to words pairs (a source word and a target word) in the ranking function (Kraaij *et al.*, 2003; Hiemstra and de Jong, 1999b). Kraaij et al. (Kraaij *et al.*, 2003) compared several models for CLIR on the CLEF-2000, 2001 and 2002 test collections. They showed that CLIR models based on mapping a query language model onto a document language model, or mapping a document language model onto a query language model, significantly outperformed CLIR models that do not use translation probabilities. They used probabilistic translation models trained on parallel texts mined from the Web. Hiemstra and de Jong (Hiemstra and de Jong, 1999b) combined the LM and translation probability from statistical dictionaries based on parallel corpora mined from the Web, instead of using one translation optimized for readability and well-formedness. They propose to improve recall by using many alternative translations. They evaluated the models on the French and Italian corpus of CLEF 2000. A performance enhancement, between 68% and 79% with respect to the monolingual baseline, was achieved.

Our main research question is to overcome the translation approaches limitations without using external resources. In parallel collections context, we propose a multilingual information model to represent the documents content (as well as queries content) in a multilingual context. It allows users to query a multilingual collection using one language and retrieve documents in a different language. The proposed structured indexing schema is combining a keywords model and a keyphrases model. For this reason, a mixture LM is used to allow the matching between keywords and between keyphrases. Our model could be useful in a wide variety of circumstances where a user is not sufficiently fluent in a document collection language to express a query in

that language. But on the other hand, the user is able to make use of the documents that are retrieved by an IRS; especially technical documents. Our Experimental evaluation is done on the MuchMore parallel collection including two languages: English and German. We compare the proposed model performances with QT performances and DT performances on the same collection presented in (Volk *et al.*, 2003).

3. Proposed CLIR Model

Our proposed CLIR model is based on the LM introduced in (Ponte and Croft, 1998). This choice is motivated by the model good performances while using few parameters. The basic idea behind this model is first to estimate a LM for each document. Then, given a query, the documents are ranked according to the likelihood that the query has been generated by the document as evidence (Zhai, 2008). Thus, for a query *Q*, the probability of a document *D* is:

$$P(Q|D) = \prod_{q_i \in Q} P(q_i|D) \quad \text{with } P(q_i|D) = \frac{tf(qi)}{l_d} \qquad [1]$$

where q_i is a query term, $tf(q_i)$ is the term frequency of q_i in the document *D* and l_d is the document length.

To avoid the zero probability when a query word does not appear in a document, we use the Dirichlet smoothing (Zhai and Lafferty, 2001). It determines the collection LM for the word q_i: $P(q_i|C)$, where C is the document collection:

$$P_{Dir}(q_i|D) = \frac{P(q_i|D) + \mu P(q_i|C)}{|D_j| + \mu} \qquad [2]$$

The value of μ, determined empirically, is set to maximize a retrieval metric (mean average precision (MAP) for example) for a set of queries and a collection of documents.

3.1. *Parallel Text Collection*

A parallel collection is a text paired with its translation into one or many languages (Resnik and Smith, 2003). It is composed of a set of document pairs in different languages that are mutual translations.

Let C be a parallel collection and L the set of languages used in C where n is the number of languages used in C. Thus, $L = \{L_1, L_2, ..., L_n\}$ where $||L|| = n$ with $n \geq 2$. The collection C is then composed by n sub-collections: $C = \{C_{L_1}, C_{L_2}, .., C_{L_n}\}$ where $L_l \in L$ and $1 \leq l \leq n$. In a parallel collection, each document is translated to *all* languages in L. The number of documents k in any sub-collection is the same: $\forall$ a sub-collection $C_{L_l} \in C$, $||C_{L_l}|| = k$ where $k \geq 1$.

3.2. *Multilingual Document Indexing*

A multilingual document in the collection C is the merging of all k paired documents. Given a multilingual document $D_j \in C$, $D_j = \{D_{j,L_1}, D_{j,L_2}, ..., D_{j,L_n}\}$ where D_{j,L_l} is the sub-document of D_j corresponding to the language L_l and $||D_j|| = n$. The document index is a logical view where a document in a collection is represented through a set of keywords. Specifically, given a text collection, a vocabulary V is created and used by the IRS to represent the documents content. The parallel collection vocabulary V is composed of two sub-vocabularies: the keywords vocabulary V_T and the keyphrases vocabulary V_K: $V = V_T \cup V_K$.

3.2.1. *Multilingual Document Indexing with Keywords*

The keywords vocabulary V_T consists of n sub-vocabularies corresponding to each specific language. Thus, given any language $L_l \in L$, a sub-vocabulary $V_{T\,L_l}$ is created and used to index any document D_{j,L_l}: $V_T = \bigcup_{i=1}^{n} V_{T\,L_i}$. Given $I_{T\,j}$ the keywords index of a multilingual document $D_j \in C$, $I_{T\,j}$ is composed of n sub-indexes: $I_{T\,j} = \bigcup_{i=1}^{n} I_{T_J,L_i}$. On the other hand, the size of a language vocabulary is different from language to another. Hence, the sizes of sub-indexes are different.

3.2.2. *Multilingual Document Indexing with Keyphrases*

The keyphrases vocabulary V_K is composed by n sub-vocabularies corresponding to each specific language. Thus, given any language $L_l \in L$, a sub-vocabulary $V_{K\,L_l}$ is created and used to index any document D_{j,L_l}: $V_K = \bigcup_{i=1}^{n} V_{K\,L_i}$. Given $I_{K\,j}$ the keyphrases index of a multilingual document $D_j \in C$, $I_{K\,j}$ is composed by n sub-indexes: $I_{K\,j} = \bigcup_{i=1}^{n} I_{K_J,L_i}$. As in the case of keywords, the size of a language keyphrases vocabulary is different from one language to another. Hence, the sizes of keyphrases sub-indexes are different.

3.3. *Structured Multilingual Documents Indexing and Matching*

Our proposed multilingual document model can be estimated using two models: keywords model (M_T) and keyphrases model (M_K). The document D_j is then represented by a structured index I_{D_j}. This document index is a sequence of w keywords and m keyphrases:

$I_{D_j} = [[[t_{j,1}, ..., t_{j,w}], [K_{j,1}, ..., K_{j,m}]]_{L_1}, ..., [[t_{j,1}, ..., t_{j,w}], [K_{j,1}, ..., K_{j,m}]]_{L_n}]$ with $t_{j,i} \in V_T$ and $K_{j,i} \in V_K$.

A query Q index is a structured index combining a sequence of r keywords and s keyphrases:

Q_q = $[[t_{q,1}, t_{q,2}, ..., t_{q,r}], [K_{q,1}, K_{q,2}..., K_{q,s}]]_{L_l}$ with $t_{q,i} \in V_T$ and $K_{q,i} \in V_K$ and the query is in the language L_l.

Although the proposed model is based on the assumption that the Keyphrases are important to capture the dependencies between words and thus find the relevant documents. It is also based on the assumption of independence between the words, effective to treat keywords. To consider the matching between keywords and the matching between keyphrases, we assume that the document model can be estimated using two models: the keyword model M_T and the keyphrases model M_K.

When combining several elements in a model, it is important to assign the appropriate weight to each part of the model. Indeed, we consider the matching between the keywords and the matching between the keyphrases as the probability that a query is generated for a given document and can be expressed as follows:

$$P(Q|D) = \alpha \prod_{i=1}^{n} P_{Dir}(t_{q,i}|D_T) + (1-\alpha) \prod_{j=1}^{m} P_{Dir}(K_{q,j}|D_K) \quad [3]$$

$P_{Dir}(t_{q,i}|D)$ and $P_{Dir}(K_{q,j}|D)$ are the corresponding probabilities for keywords and keyphrases. α is a wheighting parameter that controls the importance of the two different models, $\alpha \in [0, 1]$.

A query in the source language is matching the sub-index in the same language. After the matching process, sub-documents in different languages than the query language are returned as results. In the CLIR context, a query Q is expressed using only one language source L_s. The query index is then represented by I_Q. A CLIR system returns a set of documents $C_Q \in C$ relevant to Q using a matching function (D_j, Q) where $D_j \in C_Q$. The matching function is used to calculate the score of the document D_j for the given query Q.

4. Experimental Setup

In this section, we described the parallel collection and the experimental setup we used.

4.1. *MuchMore Parallel Collection*

To evaluate our CLIR model, we used MuchMore parallel collection (**?**). The collection contains English documents and their direct German translations. These documents have identical contents but in two languages. Documents are from 41 medical journals, each constituting a homogeneous medical sub-domain (e.g. Neurology, Radiology, etc.). Each document consists of a title, an abstract and a set of keywords. For the CLIR propose, the collection contains 25 topics. Each topic is represented by two identical queries: one English query and one German query. Each query is provided with a set of relevant documents varying from 7 relevant documents to 104 relevant

Table 1. *Parallel Collection Statistics (MuchMore Collection)*

	German Dataset	English Dataset
#Total Documents	7808	7823
# Number of terms	547696	639788
# Number of unique terms	80502	25304
#Terms by document	70	81
#Query	25	25
#Queries terms	117	132
#Terms by query	4.68	5.28

documents. The number of English relevant documents for English is 500 while it's 959 for the German relevant documents. The reason is that two different teams participated in judging the relevance (Volk *et al.*, 2002). The collection statistics are given in Table 1.

4.2. *Keyphrases Extraction*

Our approach to extract keyphrases is based on two phases:

1) We conducted a linguistic analysis with a tagger to generate a tagged collection. Each word is associated to a tag corresponding to the syntactic category of the word, for example: noun, adjective, preposition, etc. To tag the parallel collection, we used TreeTagger [4]. It is a tool for annotating text with part-of-speech and lemma information. The TreeTagger has been used to tag German and English languages.

2) Then, we used the tagged collection to extract a set of keyphrases by the identification of syntactic patterns (Bechikh-Ali *et al.*, 2015). Given S the collection lexicon, a pattern is a syntactic rule of the form:

$$X := Y_1 Y_2 Y_k ... Y_{k+1} Y_n \text{ where } Y_i \in \text{S and X is a keyphrase.}$$

For examples, *"arthroscopic treatment"* is extracted using the syntactical rule *"Adjective Noun"*. The keyphrase *"ligament injury"* is extracted using the syntactical rule *"Noun Noun"*.

Table 2 presents statistics about extracted keyphrases from documents and queries in both English and German languages.

4. http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

Table 2. *Keyphrases Extraction Statistics (in two languages)*

	German Dataset	English Dataset
#Keyphrases in documents	162125	225571
#Keyphrases by document	16.15	28.83
#Keyphrases in queries	27	48
#Keyphrases by query	1.08	1.92

4.3. *Terrier IR System and Evaluation Measures*

To implement our CLIR model, we used the Terrier IR system (Ounis *et al.*, 2005). The preprocessing includes lemmatisation with Treetagger and removing stop words. It implements various retrieval models. To investigate the effectiveness of the multilingual indexing, we implemented three different runs:

– Run_1 is implementing the M_T model indexing keywords only.

– Run_2 is implementing the M_T and the M_K models combining keywords and keyphrases but in the same documents/queries indexes. In this run, the index structure is not used.

– Run_3 implements our CLIR model introduced in section 3 where we use separate indexes one for keywords and another for keyphrases.

Run_1 and Run_2 are evaluated using three models: Tf.Idf (Salton *et al.*, 1975), BM25 (Robertson and Walker, 1994) and LM (Ponte and Croft, 1998). Run_3 is evaluated using a LM based on the matching function described in sub-section 3.3. The model BM25F (Robertson *et al.*, 2004) was used in this run instead of BM25 because it is able to evaluate a structured index considering the weighting and the matching of sub-indexes separately.

Results are evaluated using the the mean average precision (MAP) and precision measures at low recall considering only the k-top documents returned by the system: precision at 5 documents (P@5) and precision at 10 documents (P@10).

5. Evaluation Results and Discussion

Our CLIR system can perform multilingual indexing as well as monolingual indexing. To investigate the effectiveness of our model, we use the monolingual result as baseline. Table 3 shows the English and the German monolingual best results performances with the optimal parameter: the model Tf.Idf and BM25 use pivoted document lenght normalisation with a slope paramter b set by default to 0.20 for Tf.Idf and 0.75 for BM25. The LM use the parameter μ and it is set to 4500. Tf.Idf model provided the best results for the English monolingual run. For the German monolingual run, BM25 model provided the best results.

Table 3. *English Monolingual and German Monolingual Results using 25 queries*

	English Monolingual			**German Monolingual**		
Measures	Tf.Idf	BM25	LM	Tf.Idf	BM25	LM
P@5	**0.5840**	0.5760	0.5360	0.5840	**0.5920**	0.5840
P@10	**0.4920**	0.4920	0.4360	0.5080	**0.5080**	0.4800
MAP	**0.3515**	0.3511	0.3204	0.2489	**0.2530**	0.2482

Table 4. *Querying English Collection using 25 German Queries*

	Run$_1$			**Run**$_2$			**Run**$_3$	
Measures	Tf.Idf	BM25	LM	Tf.Idf	BM25	LM	BM25F (α=0.99)	LM (α=0.98)
P@5	0.5600	**0.5680**	0.5440	0.5600	0.5680	0.5440	0.4560	0.4960
P@10	**0.4920**	0.4880	0.4640	0.4640	0.4600	0.4480	0.4080	0.4120
MAP	0.2096	0.2132	**0.2139**	0.1995	0.2033	0.2069	0.1784	0.1845

5.1. *German-English CLIR*

Performance results of querying the English collection with German queries are presented in Table 4. For the Run3, we tuned the parameter α by the use of a grid search. We selected the α values that maximizes the MAP. Run_1 yields the best results compared to Run_2 and Run_3. One reason for this is the small queries size. Indeed, the average extracted keyphrases number per query is 1.08. In addition, four queries include 0 keyphrases. Comparing Run_2 and Run_3 results, we can observe that using a unified index is performing better that using the structured index. Comparing Run_3 BM25F results and Run_3 LM results, we observe that using the structured index in the case of LM model is performing better than the case of BM25F model.

5.2. *English-German CLIR*

Table 5 shows the performances of querying the German collection using English queries. The α values allowing the best MAP performances are used after tuning α by the use of a grid search. The CLIR results performances outperform the monolingual results performances. Indeed, querying a German collection with an English query gives better results than using a German query.

Querying German Collection using 25 English Queries

Table 5. *Querying German Collection using 25 English Queries*

	Run$_1$			**Run**$_2$			**Run**$_3$	
Measures	Tf.Idf	BM25	LM	Tf.Idf	BM25	LM	BM25F (α=0.99)	LM (α=0.5)
P@5	0.7200	0.7120	0.6320	**0.7440**	0.7280	0.6720	0.6240	0.5760
P@10	0.6080	0.6080	0.5480	0.6080	**0.6160**	0.5880	0.5480	0.5440
MAP	0.3126	0.3147	0.2899	0.3204	**0.3232**	0.3109	0.2854	0.2735

Combining keywords with keyphrases performs better than using only keywords. Indeed, the best results are given by Run_2 for all evaluation measures. The BM25 achieves better performances than Tf.Idf and LM models. Comparing BM25 results of Run_2 with the German monolingual results, we observe an improvement of MAP and P@10 measure with respectively 27.74% and 21.25%.

5.3. *Discussion*

To analyse the monolingual results, we processed a deep investigation on the Muchmore collection. We observed that not all the documents are translated. For example, the English document number 1449 and the German document number 1434 have no translation.

In (Volk *et al.*, 2003), QT approach and IT (semantic information combination from different resources) approach results applied to the MuchMore collection are presented. Authors limited their evaluation to the use of German queries to retrieve English documents. Comparing these results to our model results presented in Table 4, our proposed model outperformed both approaches. Indeed, our model performances exceed the QT approach performances for Run_1, Run_2 and Run_3 by respectively 55%, 50% and 33.6% for MAP and 67%, 59% and 41.1% for P@10. Run_1, Run_2 and Run_3. MAP performances exceed by respectively 20.6%, 16.6% and 4% the IT approach MAP performances and by 80.9% and 70.6% the IT approach P@10 performances. Nonetheless, we can not draw detailed conclusions because of the collection characteristics. Indeed, given the short queries size and the medical topic of the collection, few keyphrases are extracted from queries and some technical keyphrases are not extracted because they are not correctly identified by the POS tagger. That is, more experiments are needed to conclude about the efficiency of our model.

6. Conclusion

In this paper, we proposed a CLIR model avoiding translation and external resources. Our model uses a structured index to represent multilingual documents and queries. We evaluated our model on the MuchMore parallel collection (English language and German language). In the context of a parallel collection, using an English query to retrieve German documents is performing better than using a German query. We compared the proposed model performances with QT approach performances and DT approach performances. We conclude that our proposed model is performing better that both approaches. In the future, we plan to investigate more in-deep weighting schema of our model and apply our model to a larger parallel collection involving more than two languages with longer queries.

7. References

Arampatzis A., Tsoris T., Koster C. H. A., van der Weide T. P., "Phrase-Based Information Retrieval", *Inf. Process. Manage.*, vol. 34, n° 6, p. 693-707, 1998.

Bechikh-Ali C., Wang R., Haddad H., "A Two-Level Keyphrase Extraction Approach", *The 16th International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt*, p. 390-401, 2015.

Chew P. A., Abdelali A., "Benefits of the 'Massively Parallel Rosetta Stone': Cross-Language Information Retrieval with over 30 Languages", *the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic*, p. 872-879, 2007.

Costa-jussà M Banchs R. E., "Is there Hope for Interlingua methods? A CLIR comparison experiment between Interlingua and Query Translation", *Research in Computing Science*, vol. 74, p. 81-87, 2014.

Gao J., Nie J., Zhou M., "Statistical query translation models for cross-language information retrieval", *ACM Trans. Asian Lang. Inf. Process.*, vol. 5, n° 4, p. 323-359, 2006.

Gey F. C., "How Similar are Chinese and Japanese for Cross-Language Information Retrieval?", *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-5, National Center of Sciences, Tokyo, Japan, December 6-9, 2005*, 2005.

Haddad H., Bechikh-Ali C., "Performance of Turkish Information Retrieval: Evaluating the Impact of Linguistic Parameters and Compound Nouns", *The 15th International Conference on Intelligent Text Processing and Computational Linguistics, Kathmandu, Nepal*, p. 381-391, 2014.

Hieber F., Riezler S., "Bag-of-Words Forced Decoding for Cross-Lingual Information Retrieval", *The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Colorado, USA*, p. 1172-1182, 2015.

Hiemstra D., de Jong F., "Cross-language information retrieval in Twenty-One: Using one, some or all possible translations?.", *In the 14th Twente Workshop on Language Technology, Twente, Netherlands*, p. 19-25, 1999a.

Hiemstra D., de Jong F., "Disambiguation Strategies for Cross-Language Information Retrieval", *The Third European Conference on Research and Advanced Technology for Digital Libraries, London, UK*, ECDL '99, p. 274-293, 1999b.

Hu R., Chen W., Bai P., Lu Y., Chen Z., Yang Q., "Web Query Translation via Web Log Mining", *The 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore*, p. 749-750, 2008.

Kishida K., "Technical issues of cross-language information retrieval: a review", *Inf. Process. Manage.*, vol. 41, n° 3, p. 433-455, 2005.

Kraaij W., Nie J., Simard M., "Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval", *Computational Linguistics*, vol. 29, n° 3, p. 381-419, 2003.

Nguyen D., Overwijk A., Hauff C., Trieschnigg D., Hiemstra D., de Jong F., "WikiTranslate: Query Translation for Cross-Lingual Information Retrieval Using Only Wikipedia", *The 9th Workshop of the Cross-Language Evaluation Forum, Evaluating Systems for Multilingual and Multimodal Information Access, Aarhus, Denmark*, p. 58-65, 2008.

Oard D. W., Diekema A. R., "Cross-language information retrieval", *Annual review of information science and technology*, vol. 33, p. 223-256, 1998.

Ounis I., Amati G., Plachouras V., He B., Macdonald C., Johnson D., "Terrier Information Retrieval Platform", *The 27th European Conference on Iinformation Retrieval Research, Santiago de Compostela, Spain*, p. 517-519, 2005.

Peters C., Braschler M., Clough P. D., *Multilingual Information Retrieval - From Research To Practice*, Springer, 2012.

Ponte J. M., Croft W. B., "A Language Modeling Approach to Information Retrieval", *The 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia*, p. 275-281, 1998.

Resnik P., Smith N. A., "The Web as a Parallel Corpus", *Computational Linguistics*, vol. 29, n° 3, p. 349-380, 2003.

Robertson S. E., Walker S., "Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval", *The 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland*, p. 232-241, 1994.

Robertson S. E., Zaragoza H., Taylor M. J., "Simple BM25 extension to multiple weighted fields", *The thirteenth ACM international conference on Information and knowledge management, Washington, DC, USA*, p. 42-49, 2004.

Salton G., Yang C. S., Yu C. T., "A theory of term importance in automatic text analysis", *Journal of the American Society for Information Science*, vol. 26, n° 1, p. 33-44, 1975.

Ture F., Elsayed T., Lin J., "No Free Lunch: Brute Force vs. Locality-sensitive Hashing for Cross-lingual Pairwise Similarity", *The 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China*, p. 943-952, 2011.

Volk M., Ripplinger B., Vintar S., Buitelaar P., Raileanu D., Sacaleanu B., "Semantic annotation for concept-based cross-language medical information retrieval", *I. J. Medical Informatics*, vol. 67, n° 1-3, p. 97-112, 2002.

Volk M., Vintar S., Buitelaar P., "Ontologies in Cross-Language Information Retrieval", *2nd Conference on Professional Knowledge Management, Luzern, Switzerland*, p. 43-50, 2003.

Zhai C., “Statistical Language Models for Information Retrieval: A Critical Review”, *Foundations and Trends in Information Retrieval*, vol. 2, n° 3, p. 137-213, 2008.

Zhai C., Lafferty J. D., “A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval”, *The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Louisiana, USA*, p. 334-342, 2001.

Zhou D., Truran M., Brailsford T. J., Wade V., Ashman H., “Translation techniques in cross-language information retrieval”, *ACM Comput. Surv.*, vol. 45, n° 1, p. 1, 2012.