Noname manuscript No.
(will be inserted by the editor)

An Axiomatic Approach to Corpus-Based Cross-Language Information Retrieval

Razieh Rahimi · Ali Montazeralghaem · Azadeh Shakery

Received: date / Accepted: date

Abstract A major challenge in Cross-Language Information Retrieval (CLIR) is the adoption of translation knowledge in retrieval models, as it affects term weighting which is known to highly impact the retrieval performance. Despite its importance, how different approaches for integration of translation knowledge into retrieval models relatively perform has not been analytically examined. In this paper, we present an analytical investigation of using translation knowledge in CLIR. In particular, by adopting the axiomatic analysis framework, we formulate impacts of using translation knowledge on document ranking as constraints that any cross-language retrieval model should satisfy. We then consider state-of-the-art CLIR methods and check whether they satisfy these constraints. Our study shows that none of the existing methods satisfies all constraints. Based on the defined constraints, we propose the hierarchical query modeling method for CLIR which satisfies more constraints and achieves a higher CLIR performance, compared to the existing methods.

Keywords Axiomatic analysis · cross-language information retrieval · probabilistic structured query

R. Rahimi
Center for Intelligent Information Retrieval, University of Massachusetts Amherst
E-mail: rahimi@cs.umass.edu

Ali Montazeralghaem,
Center for Intelligent Information Retrieval, University of Massachusetts Amherst
E-mail: montazer@cs.umass.edu

A. Shakery
School of Electrical and Computer Engineering, College of Engineering, University of Tehran
School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
E-mail: shakery@ut.ac.ir

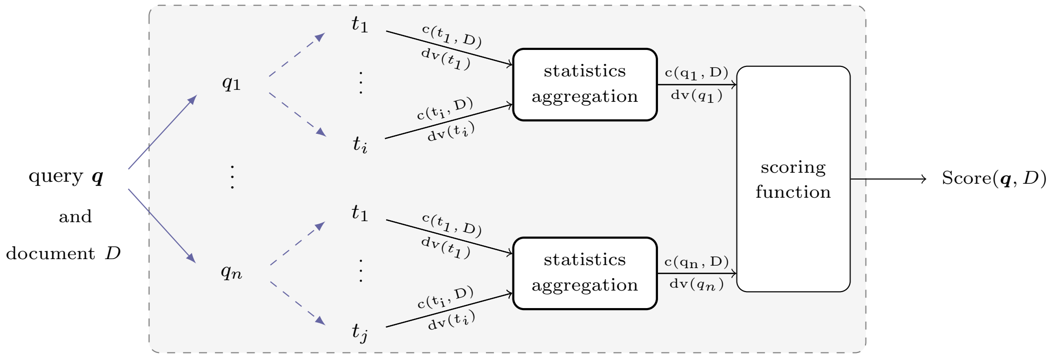
1 Introduction

Performance optimization of systems for information retrieval is a long standing yet challenging problem, which has attracted a great deal of theoretical and experimental attention in the literature. Among different types of search tasks in information retrieval systems, evaluation of models for Cross-Language Information Retrieval (CLIR) has been less thoroughly studied. CLIR systems allow users to formulate queries in one language, usually their native language, in order to seek information in another language. The general and promising approach is to use some sort of translation resource for crossing the language barrier between the query and the documents. Machine-readable bilingual dictionaries do not provide sufficient coverage for CLIR due to out of vocabulary words and neologisms. To compensate this deficiency, CLIR approaches tend to use statistical translation models learned from other translation resources, such as bilingual corpora, to achieve acceptable performance. Using translation models is thus essential to perform effective CLIR. Herein, we focus on corpus-based CLIR using translation models learned from aligned corpora.

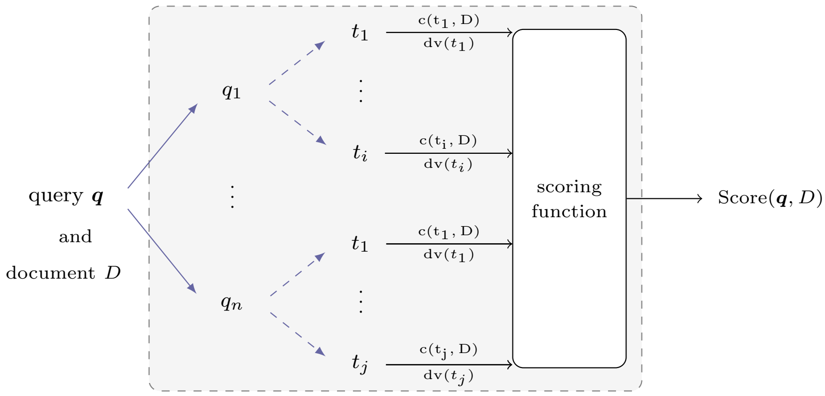
Different approaches have been proposed for the integration of translation models into retrieval systems. In keeping with the dominant approach for translation-based CLIR, we focus on the query translation approach (Nie 2010). This type of approach for CLIR has been developed along two main lines in terms of translation model integration: (1) representing a query based on weighting each query term by aggregating its translations' statistics; and (2) representing a query by weighting translations in the target language (documents' language). These two categories are depicted in Fig. 1. The *probabilistic structured query* (PSQ) method (Darwish and Oard 2003) is a sample of the first category, while cross-language information retrieval based on the language modeling framework (henceforth referred to as the LM-based method) belongs to the second category.

How to integrate translation models into retrieval functions significantly impacts the performance of CLIR. This is because retrieval functions calculate the score of a document based on retrieval heuristics, e.g., *Term Frequency* (TF) and *Document Frequency* (DF), which, in case of CLIR, are estimated based on how the translation model is integrated into the retrieval process. Both PSQ and LM-based methods, as samples of the two approaches for integration of translation models, show promising results, but none consistently outperforms the other. Therefore, to improve the performance of CLIR, there is a fundamental need for analytical investigation of the CLIR models, primarily based on the influential factor of integrating translation models into retrieval functions. However, to the best of our knowledge, the impacts of translation models in cross-language environments have not yet been analytically studied.

We aim to analytically investigate the CLIR models in order to improve the effectiveness of CLIR. Analytical investigation of CLIR models allows us to gain more insights into the impacts of translation models on retrieval heuristics. More specifically, our analytical study shows that given a translation



(a) The PSQ method.



(b) The LM-based method.

Fig. 1 Existing approaches for integration of translations in cross-language information retrieval.

model, what estimations of term frequencies or discrimination values provide reasonable document ranking in CLIR. For this purpose, we follow the axiomatic analysis approach (Fang et al 2004), where the desirable impacts of retrieval heuristics are formulated as constraints that any reasonable retrieval model should satisfy. This analysis has been extensively used to diagnose different aspects of monolingual retrieval models, and has led to significant improvements in the performance of monolingual retrieval (see Fang et al (2011); Lv and Zhai (2011); Wu and Fang (2012); Clinchant and Gaussier (2013); Pal et al (2015); Montazeralghaem et al (2017) among others).

Following the axiomatic analysis, we aim to define specific constraints on the scoring functions of CLIR models so that they provide reasonable rankings of documents. The first step in this task is to define constraints to regulate how two extremely important factors of term frequency and term discrimination values should be estimated in the CLIR setting and used in scoring functions. Utilization of translation resources into CLIR models makes this task particularly challenging. We propose three constraints that a reasonable CLIR model should satisfy; two constraints on translation coverage, and one constraint on discrimination values of translations. The defined constraints show important aspects of integrating translation models into scoring functions of CLIR models.

Checking the defined constraints on the PSQ and LM-based methods as representatives of the two different approaches for integration of translation models, demonstrates two points: (1) State-of-the-art CLIR models and subsequently their corresponding approach for integration of translation models do not satisfy all constraints unconditionally. In case of weighting each query term by aggregating its translations' statistics (such as the PSQ method), documents that cover more distinct or more discriminative translations of a query term may not be distinguished. On the other hand, in case of representing a query by weighting its translations (such as the LM-based model), documents covering translations of more distinct query terms may not be distinguished. This point justifies the reason why none of the CLIR methods consistently outperforms the other. (2) A CLIR model to perform optimally should simultaneously consider the statistics used in both translation-integration approaches; statistics of translations and statistics of query terms by aggregating their translations's statistics. This point shows the path for improving the performance of existing CLIR models or proposing a new CLIR model.

Based on our findings, we improve the probabilistic structured query method step by step so that it satisfies one more constraint at each step. We first propose an aggregation function for the frequency of a query term which is also sensitive to the number of translations of the query term occurring in a document. This improvement is particularly suited to the case that one translation of a query term has a general use in the target language, then a document which covers other more specific translations of the query term can be relevant to the query with a higher probability. Second, we propose a function to estimate discrimination values of query terms. The proposed function ranks documents based on hierarchical estimation of discrimination values as follows; when two translations have different discrimination values, the document covering the more discriminating translation is ranked higher. Otherwise, the document whose covered translations of a query term have a higher aggregated discrimination value is considered to be more relevant. We subsequently combine all modifications into one final variant and propose the Hierarchical Query Modeling method for CLIR, which satisfies two more constraints compared to the original version of the PSQ method.

Finally, we empirically evaluate the impacts of the proposed constraints on the performance of CLIR. First, we compare the performance of PSQ and LM-based methods, and show that the results of constraint analysis on the CLIR methods correlate with their performance on test datasets. Second, the stepwise improvement enables us to demonstrate the impact of satisfying each constraint on the performance of CLIR. Our empirical evaluation on datasets in various languages reveals that the proposed variants of the probabilistic structured method improves the performance of CLIR.

The remainder of this paper is organized as follows. In Section 2, we review the previous work. Then, we present our proposed constraints for CLIR models in Section 3. Following, analysis of the CLIR models with respect to the derived constraints are reported in Section 4. In Section 5, the improved model for CLIR is described, and defined constraints are checked against the new proposed model. Experimental settings are described in Section 6 and the results of our empirical evaluations are reported and discussed in Section 7. Finally, the paper is concluded in Section 8.

2 Related Work

Axiomatic analysis, introduced by Fang et al. (2004; 2005), is based on formal constraints that any reasonable retrieval model should satisfy. Several constraints have been defined for different factors impacting the performance of monolingual information retrieval. Fang et al. (2004; 2011) defined basic constraints on the impacts of term frequency, document frequency, and document length on the scores of documents. Following this line, Lv and Zhai (2011) defined constraints regarding the normalization of term frequency to avoid over-penalization of very long documents.

In addition to constraints on basic heuristics used in retrieval models, axiomatic analysis has been applied to different components of information retrieval systems. For example, Clinchant and Gaussier (2011; 2013) investigate adoption of pseudo-relevance feedback (PRF) in monolingual information retrieval models. Montazeralghaem et al. (2016) proposed two constraints for pseudo-relevance feedback where one is about semantic similarity between feedback terms and query terms and the other is about the distribution of terms in feedback documents. Pal et al. (2015) introduced a constraint for PRF models in the divergence from randomness framework, which is about relevance scores of feedback documents in PRF models. Similar to that work, Ariannezhad et al. (2017b) proposed a constraint about the effect of feedback terms' weights on relevance scores of feedback documents.

Similar to this line of research, some studies define constraints to regulate query/document expansion to enable matching of semantically related terms in addition to exact term matching (Fang and Zhai 2006; Fang 2008; Zheng and Fang 2010; Wu and Fang 2012; Karimzadehgan and Zhai 2012). In this line, constraints on the principles of matching related terms are defined in (Fang and Zhai 2006; Fang 2008).

Considering other aspects that impact the performance of retrieval, Tao and Zhai (2007) define constraints on the proximity of matched query terms in documents and propose different proximity measures. Following this work, Montazeralghaem et al. (2017) proposed three constraints for PRF models on the proximity of feedback terms to query terms in feedback documents.

Na et al. (2008) study normalization of term frequency in documents that cover multiple topics. Gerani et al. (2012) define constraints on document scoring by linear combination of different ranking aspects.

Information retrieval for verbose queries has also been examined using the axiomatic analysis framework. Lv (2015) proposed a constraint to consider the relation between query length and the term frequency decay speed. Ariannezhad et al. (2017a) also proposed a constraint to model relation between query length and the term discrimination of terms for verbose queries.

The axiomatic analysis framework has also been used to define constraints on evaluation measures for information retrieval systems in (Amigó et al 2013; Busin and Mizzaro 2013). All these studies focus on investigating monolingual retrieval models.

Although there is a substantial body of research on analytical study of monolingual retrieval models, the corresponding literature on cross-language retrieval models is very thin. Indeed, to the best of our knowledge, the only relevant studies are (Kern et al 2009; Karimzadehgan and Zhai 2012). But, none of these studies fulfills our goal in this article, which is to define formal constraints specific to any reasonable CLIR model. In particular, Kern et al. (2009) adopted the corrected BM25 retrieval model, proposed in (Fang and Zhai 2005) for monolingual retrieval through axiomatic analysis, for document ranking in CLIR. This work, which employs the optimal monolingual retrieval function, is thus different from ours on deriving constraints on any CLIR model. The proposed constraints in (Karimzadehgan and Zhai 2012) are to regulate the estimation of relations between words in one language. These monolingual word relationships, referred to as translation models, are then used to estimate more accurate language models for documents to improve the performance of monolingual information retrieval.

Li and Gaussier (2012) extend the information-based model for monolingual information retrieval to the cross-lingual setting. The proposed retrieval model is a dictionary-based model for CLIR, which assumes uniform weights for all translations of a term. They also propose one constraint on CLIR models; consider a single term query, and two equal length documents where one document has one occurrence of k different translations of the query term, and the other document has k occurrences of one translation. Also assume that all translations of the query term have equal discrimination values, and are all equally good translations. The two documents then should have the same retrieval score with respect to that query. Li et al. Li et al (2018) extend this CLIR constraint to queries with two terms in the similar settings. Translations in human-constructed dictionaries differ from those in probabilistic translation models learned from bilingual corpora (Nie 2010). Unlike probabilistic translation models, dictionaries do not contain noisy translations nor contain words that are related to or co-occur with the translations of a term. Therefore, the impacts of translation coverage in case of using human-constructed dictionaries are different than those of using automatically-built probabilistic dictionaries.

Cross-language information retrieval. The task of CLIR is to score documents with respect to a query in another language than that of the documents. Due to the different languages of queries and documents, some sort of processing is needed to match document terms with query terms. Cross-language information retrieval between similar language pairs (such as Italian-French and Chinese-Japanese (Savoy 2005)) can be performed without any direct translation (Buckley et al 2000; He et al 2003; Mcnamee and Mayfield 2004). However, the most general approach for this task is to use translation resources.

Translation knowledge is used in CLIR to make a comparable representation of both queries and documents. Building comparable representations of queries and documents can be done using different strategies; by representing both queries and documents either in the query language space, or in the document language space, in an intermediate language space or in low-dimensional vectors (Kraaij and de Jong 2004; Sorg and Cimiano 2012; Vulić and Moens 2015). The low-dimensional vectors for the CLIR task, proposed by Vulić and Moens (2015), can be considered as an intermediate language space for queries and documents, because word embedding is used for representation of words in documents and queries which are in two languages. Each strategy has its own advantages and limitations. The goal of our work is to study how to effectively use translation knowledge to build representations in other languages, and we focus on query translation strategy.

As mentioned in the Introduction section, different approaches for using translation knowledge in retrieval models can be categorized into two groups. The first category of approaches adopts the idea of translation models in monolingual information retrieval, proposed in (Berger and Lafferty 1999), to CLIR. The cross-lingual models proposed in (Xu et al 2001; Lavrenko et al 2002; Kraaij et al 2003) belong to the first category. More specifically, Xu et al. (2001) used a general collection in the query language for smoothing the new estimated language models for documents, while Kraaij et al. (2003) smoothed the document language models using the reference language model of document collection in the target language. In our experiments, we follow the latter choice for smoothing document language models. On the other hand, the probabilistic structured query model proposed in (Darwish and Oard 2003) belongs to the second category of approaches, where each query term is weighted using an aggregation function on statistics of its translations.

Empirical evaluation of cross-language retrieval models are studied in (Oard and Wang 2001). They empirically compare the performance of Pirkola's structured queries with balanced translation for English-Chinese information retrieval, and show that the Pirkola's structured query method outperforms balanced translation.

3 Constraints on CLIR Models

The focus here is to study the corpus-based CLIR models that try to optimize the performance of retrieval based on translation probabilities learned from aligned corpora. Among different architectures for CLIR models, we study the query translation approach where queries are represented in the document representation space. Therefore, a translation model P of the form $p(w|u)$, which indicates the probability of translating term u in the query language to term w in the document language, is given. In addition, translation probabilities are normalized in the query translation direction. The goal is to define constraints concerning frequencies and discrimination values of query terms that any reasonable CLIR model based on the query translation strategy should satisfy. Before proceeding to the formal representation of constraints, we introduce some notation. The frequency of a term w in document D is denoted by $c(w, D)$. The $\mathrm{df}(w)$ and $\mathrm{dv}(w)$ represent document frequency and discrimination value of term w, respectively. We denote the jth translation of query term q^i by t^i_j. Constraints are named following the CLIR constraints defined by Rahimi et al. (2014).

3.1 Constraints on Translation Coverage

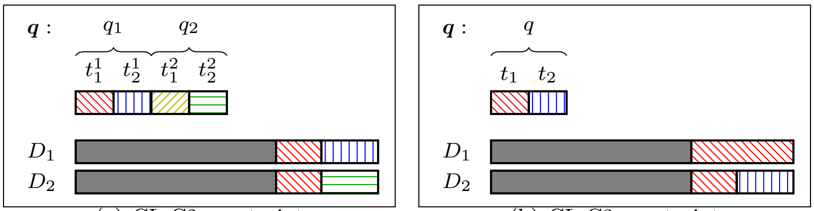
The goal in this section is to define constraints that concern the coverage of distinct query or translation terms, similar to the purpose behind the TFC3 constraint on models for monolingual information retrieval (Fang et al 2011). Generally, we aim to favor documents firstly in terms of covering more distinct query terms and then based on containing more translation alternatives.

CL-C2 constraint. The first CLIR constraint regarding translation coverage is about the coverage of translations of distinct query terms. For illustration, consider the example in Fig. 2(a), where two translation terms t^1_1 and t^1_2 occur in document D_1 with the same total number as the occurrences of t^1_1 and t^2_2 in document D_2. However, D_1 covers only the translations of one query term q_1, while D_2 covers the translations of both query terms q_1 and q_2. Assuming that t^1_2 and t^2_2 have the same discrimination value, D_2 should get a higher score since it covers translations of more distinct query terms. The formal statement of CL-C2 is as follows:

Suppose we have a two-term query $\boldsymbol{q} = q_1 q_2$ and one document D. Translation model includes translations t^1_i and t^1_j for query term q_1 and translation t^2_k for q_2, where $p(t^1_j|q_1) = p(t^2_k|q_2)$. Also assume that among all translations of query terms, document D covers only one translation t^1_i of query term q_1. If two translations t^1_j and t^2_k have the same discrimination value, i.e., $\mathrm{dv}(t^1_j) = \mathrm{dv}(t^2_k)$, then $s(q, D_1) < s(q, D_2)$, where $D_1 = D \cup \{t^1_j\}$ and $D_2 = D \cup \{t^2_k\}$.

CL-C3 Constraint. This constraint is about the coverage of different translation alternatives of a query term. As an illustration, consider the example in Fig. 2(b). Two documents D_1 and D_2 have the same total occurrences of t_1 and t_2, which are translations of query term q with equal probabilities. However, D_2 covers two distinct translations of query term q, while D_1 covers only one translation of q. Assuming that t_1 and t_2 have the same discrimination value, D_2 should get a higher score with respect to query $\boldsymbol{q}$.

Suppose we have a one-term query $\boldsymbol{q} = q$ and two documents D_1 and D_2 with the same length. Translation model includes translations t_1 and t_2



(a) CL-C2 constraint (b) CL-C3 constraint

Fig. 2 Examples of CLIR constraints on translation coverage.

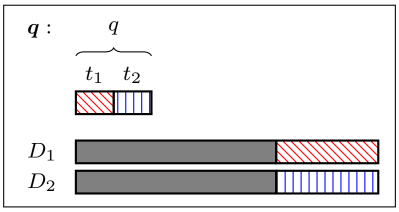


Fig. 3 Example of CL-C4 constraint regarding query term discrimination.

for query term q with $p(t_1|q) = p(t_2|q)$ probabilities. Also assume that two documents have equal frequencies of translation terms, $\mathrm{c}(t_1, D_1) + \mathrm{c}(t_2, D_1) = \mathrm{c}(t_1, D_2) + \mathrm{c}(t_2, D_2)$ where $\mathrm{c}(t_2, D_2) > 0$ and $\mathrm{c}(t_1, D_2) > 0$, but $\mathrm{c}(t_2, D_1) = 0$. If two translations have the same discrimination value, i.e., $\mathrm{dv}(t_1) = \mathrm{dv}(t_2)$, then $s(q, D_1) < s(q, D_2)$.

This constraint can be derived based on the concavity of scoring functions. The CL-C3 constraint is merely a tiebreaker rule when t_1 and t_2 are synonyms. In the setting of our study, where translation models contain related words, satisfying this constraint can have an effect similar to that of query expansion.

3.2 Constraint on Term Discrimination Values

The constraint defined in this section regulates the interaction between *frequencies* and *discrimination values* of query terms.

CL-C4 constraint. The intent of this constraint is to ensure that adding a translation alternative with a higher discrimination value to a document increases the score of the document more. Fig. 3 shows an example of this constraint. In this sample, two documents D_1 and D_2 have the same total number of occurrences of two translation alternatives of query term q. When the two translation alternatives have the same translation probability, the document which has more occurrences of the more specific translation alternative should get a higher score.

The CL-C4 constraint can be formalized as follows. Suppose we have a one-term query $\boldsymbol{q} = q$ and document D. Translation model includes translations t_1 and t_2 for query term q with the same translation probability, $p(t_1|q) = p(t_2|q)$.

Also, assume that $\mathrm{dv}(t_1) > \mathrm{dv}(t_2)$. If $c(t_1, D) \leq c(t_2, D)$, $D_1 = D \cup \{t_1\}$, and $D_2 = D \cup \{t_2\}$, then one should have $s(q, D_1) > s(q, D_2)$.

This constraint states that adding a translation alternative of the query term that has a higher discrimination value increases the retrieval score of a document more. Therefore, translation alternatives with higher discrimination values are preferred to be added to a document. The $c(t_1, D) \leq c(t_2, D)$ condition, avoids over-favoring translation alternatives with high discrimination values.

4 Constraint Analysis on CLIR Models

4.1 Representative CLIR Models

In this section, we briefly describe two CLIR models which seek to optimize retrieval effectiveness through reliance on obtained translation probabilities from aligned corpora.

PSQ method. In this method (Darwish and Oard 2003), translation probabilities are considered in the computation of term frequencies and document frequencies of query terms as follows:

$$c(q_i, D) = \sum_{t \in V_T} p(t|q_i) \times c(t, D), \tag{1}$$

$$\mathrm{df}(q_i) = \sum_{t \in V_T} p(t|q_i) \times \mathrm{df}(t), \tag{2}$$

where $t \in V_T$ is a term belonging to the vocabulary set of the target language (the language of documents), $q_i \in \boldsymbol{q}$ is a query term, $p(t|q_i)$ is the probability of translating word q_i into word t, and $\mathrm{df}(.)$ represents the document frequency of a term. These estimates of term and document frequencies are then used in BM25 retrieval model to score document D with respect to query $\boldsymbol{q}$. In BM25 retrieval model for monolingual information retrieval, the score of a document with respect to a query is computed using:

$$\begin{aligned} &\mathrm{S}_{\mathrm{BM25}}(\boldsymbol{q}, D) \\ &= \sum_{q_i \in D} \Big(\mathrm{dv}(q_i) \times \frac{(k_1 + 1) \times c(q_i, D)}{k_1\big((1 - b) + b\frac{|D|}{\mathrm{avdl}}\big) + c(q_i, D)} \times \frac{(k_3 + 1) \times c(q_i, \boldsymbol{q})}{k_3 + c(q_i, \boldsymbol{q})} \Big), \end{aligned} \tag{3}$$

where k_1 and k_3 are free parameters, and avdl denotes the average length of documents in the collection. The discrimination values of query terms can be evaluated in different ways. We follow the below estimation so that BM25 retrieval model satisfies all term-frequency constraints (Fang et al 2011).

$$\mathrm{dv}(q_i) = \ln(\frac{N + 1}{\mathrm{df}(q_i)}), \tag{4}$$

where N is the number of documents in the collection. In the following analysis, we assume that the adopted BM25 retrieval model in the PSQ method satisfies all term-frequency constraints unconditionally.

LM-based model. The KL-Divergence retrieval model provides a unified framework for translation and retrieval steps in CLIR (Nie 2010). In the query translation approach, a new language model is built for the query and documents are ranked using:

$$\mathrm{S}_{\mathrm{LM-based}}(\boldsymbol{q}, D) = \sum_{t \in V_T} p(t|\theta'_q) \log p(t|\theta_D), \tag{5}$$

$$p(t|\theta'_q) \approx \sum_{s \in V_S} p(t|s) p(s|\theta_q), \tag{6}$$

where s and t are source and target words respectively, $p(t|s)$ indicates the translation probability, θ_q represents the query language model in the source language, θ'_q denotes the new language model in the target language for the query, and θ_D shows the document language model. We assume that document language models are obtained using the Dirichlet prior smoothing method. Using this smoothing method, the scoring function of Eq. (5) becomes:

$$\mathrm{S}_{\mathrm{LM-based}}(\boldsymbol{q}, D) = \sum_{\substack{w: p(w|\theta'_q)>0 \\ c(w,D)>0}} p(w|\theta'_q) \log(1 + \frac{c(w, D)}{\mu p(w|C)}) + \log \frac{\mu}{\mu + |D|}, \tag{7}$$

where μ is the parameter of the Dirichlet prior smoothing method, and $p(.|C)$ denotes the collection language model.

In the following sections, we analyze the two CLIR models to validate whether they satisfy the defined constraints on CLIR models. Before proceeding further we note that BM25 and the language modeling framework, which underlie the mentioned methods for CLIR, satisfy all constraints related to term frequency - TFC1, TFC2, and TFC3 - and term discrimination values TDC (Fang et al 2011).

4.2 Constraint on Translation Coverage

In this section, we check whether the CLIR models satisfy the constraints on translation coverage, i.e., CL-C2 and CL-C3 constraints.

PSQ method. Applying the assumptions of CL-C2 into Eq. 1, one has $c(q_1, D_1) + c(q_2, D_1) = c(q_1, D_2) + c(q_2, D_2)$, i.e., two documents have the same total frequencies of query terms. But, document D_2 covers q_1 and q_2, while document D_1 covers only q_1. Hence, D_2 covers more distinct query terms. Document frequencies $\mathrm{df}(q_1)$ and $\mathrm{df}(q_2)$ in the PSQ method are estimated using Eq. 2 which depends on all translation alternatives for query terms. Therefore, according to the assumptions of CL-C2, we cannot compare $\mathrm{df}(q_1)$ and $\mathrm{df}(q_2)$. If $\mathrm{df}(q_1) = \mathrm{df}(q_2)$, then the PSQ method satisfies the CL-C2 constraint, because the BM25 retrieval model prefers a document covering more distinct query terms in the conditions above (Fang et al 2011).

According to the assumptions of CL-C3, one has $c(q, D_1) = c(q, D_2)$. Thus, documents D_1 and D_2 get the same score using the PSQ method which means that this does not satisfy the CL-C3 constraint.

LM-based model. To validate constraints about translation coverage on the LM-based method, we first estimate the new query language model using Eq. 6 given the assumptions of the constraints. Validation is then performed according to the behavior of the language modeling framework given the new estimated language model for the query.

Applying the assumptions of the CL-C2 constraint into Eq. 6, one has $p(t_j|\theta'_q) = p(t_k|\theta'_q)$ for the new query language model. Also, for the language models of augmented documents, one has $p(t_j|\theta_{D_1}) = p(t_k|\theta_{D_2})$ and $p(t_i|\theta_{D_1}) = p(t_i|\theta_{D_2})$ based on the assumptions of the CL-C2 constraint, since two documents D_1 and D_2 have the same length. Therefore, the scores of the two documents, which are calculated using Eq. 5 as below are equal. Thus, the LM-based model does not satisfy the CL-C2 constraint.

$$\mathrm{S}(\boldsymbol{q}, D_1) = \mathrm{S}(\boldsymbol{q}, D) + p(t_j|\theta'_q)p(t_j|\theta_{D_1}), \tag{8}$$
$$\mathrm{S}(\boldsymbol{q}, D_2) = \mathrm{S}(\boldsymbol{q}, D) + p(t_k|\theta'_q)p(t_k|\theta_{D_2}). \tag{9}$$

Applying the assumptions of the CL-C3 constraint into Eq. 6, one has $p(t_1|\theta'_q) = p(t_2|\theta'_q) > 0$ for the new query language model. Also, two documents cover different numbers of terms that have non-zero probabilities in the new query language model, while both documents have the same total occurrences of these terms. More specifically, probabilities of query terms in document language models are as follows:

$$p(t_1|\theta_{D_1}) > 0, \quad p(t_1|\theta_{D_2}) > 0,$$
$$p(t_2|\theta_{D_1}) = 0, \quad p(t_2|\theta_{D_2}) > 0.$$

In these conditions, the language modeling framework prefers document D_2 over document D_1 (according to TFC3 in (Fang et al 2011)). Therefore, LM-based model satisfies the CL-C3 constraint.

4.3 Constraint on Term Discrimination Values

In this section, we evaluate the CLIR models based on the CL-C4 constraint concerning query term discrimination.

PSQ method. The PSQ method does not satisfy the CL-C4 constraint. Based on the assumptions of the constraint, both terms t_1 and t_2 are translations of one query term and have equal translation probabilities, one has $c(q, D_1) = c(q, D_2)$ obtained using Eq. (1). Therefore, the BM25 model assigns equal scores to both documents D_1 and D_2.

LM-based model. This method satisfies the CL-C4 constraint. The reason is that the language modeling framework for monolingual information

Table 1 Summary of constraint analysis results of cross-lingual retrieval models.

	Term Coverage		Term Discrimination Values
	CL-C2	CL-C3	CL-C4
PSQ	Cond	No	No
LM-based	No	Yes	Yes
HQM	Cond	Yes	Cond

retrieval satisfies the TDC axiom. Thus, the LM-based model, operating on a new query language model in CLIR, prefers a document covering the translation alternative with a higher discrimination value.

Table 1 shows the summary of constraint satisfaction status of different CLIR models.

5 Improving the PSQ Method

In this section, we propose two modifications on the existing PSQ method so that it satisfies more of the constraints on the scoring function.

5.1 Improving the Estimation of Query Term Frequency

According to the analysis in Section 4.2, the PSQ method does not satisfy the CL-C3 constraint on preferring a document that covers more translation alternatives of a query term. To comply with the CL-C3 constraint, we modify the PSQ method as follows:

$$c'(q_i, D) = c(q_i, D) \times \log(h + \sigma), \tag{10}$$

$$c(q_i, D) = \sum_{t \in V_T} p(t|q_i) \times c(t, D), \tag{11}$$

$$h = \sum_{t \in V_T} \mathbf{1}_{p(t|q_i) \times c(t,D)}, \tag{12}$$

where frequencies of query terms are estimated by an additional factor $\log(h + \sigma)$ where h shows the number of translation alternatives with non-zero counts in the document. Parameter σ is a smoothing parameter that is designed to avoid the zero problem, and to reduce the difference between different numbers of translations covered by a document. In addition, σ allows the obtained retrieval model to still satisfy the CL-C2 constraint. We later show in the experiments that parameter σ depends on the characteristics of datasets, and by appropriately setting this parameter, the modified PSQ method satisfies both CL-C2 and CL-C3 constraints almost unconditionally. This new estimation of query term frequencies together with the estimation of term discrimination values in Eq. 2 can be adopted in the BM25 retrieval model. We refer to this retrieval function as PSQ+CL-C3.

5.2 Improving the Estimation of Term Discrimination Values

As demonstrated in Section 4.3, the original PSQ method does not satisfy the CL-C4 constraint concerning the discrimination values of translation alternatives. The reason is due to the aggregation of statistics of all translations of a query term using Eq. 2, therefore different translations of a query term cannot be distinguished in the scoring of documents. In order to overcome this deficiency, we propose to separate the impacts of translation alternatives which do and do not exist in a document as follows:

$$\mathrm{df}(q_i) = \sum_{t \in V_T} p(t|q_i) \times \mathrm{df}(t), \tag{13}$$

$$\mathrm{df}(q_i, D) = \sum_{t \in D} p(t|q_i) \times \mathrm{df}(t), \tag{14}$$

$$\mathrm{dv}(q_i, D) = \ln(\frac{N+1}{\mathrm{df}(q_i, D) + 0.5}) \times \frac{\ln(\frac{N+1}{\mathrm{df}(q_i) - \mathrm{df}(q_i, D) + 0.5})}{\ln(\frac{N+1}{\mathrm{df}(q_i) - \mathrm{df}(q_i, D) + 0.5}) + c}, \tag{15}$$

where $\mathrm{df}(q_i, D)$ aggregates the document frequencies of translations of q_i occurring in document D. The discrimination value of q_i is then estimated using both $\mathrm{df}(q_i)$ and $\mathrm{df}(q_i, D)$ in order to distinguish documents that have translations with different discrimination values. The parameter $c > 0$ is a free parameter which controls the contribution of translation alternatives not occurred in a document on the document's score. The parameter c should be set to a value higher than zero, otherwise translations not occurred in a document do not impact the discrimination value of the query term. The greater the value of parameter c, the lower the impacts of translations not occurred in a document on the discrimination value of the query term.

This new estimation of term discrimination values together with the estimation of query term frequencies in Eq. 1 can be adopted in the BM25 retrieval model. We refer to this retrieval function as PSQ+CL-C4.

5.3 Hierarchical Query Modeling

Finally, we propose to adopt the two new estimation methods for query term frequencies and term discrimination values, defined in Eqs. 12 and 15 respectively, in the BM25 retrieval model to perform cross-language information retrieval. We refer to this retrieval function as *hierarchical query modeling* (HQM), since both statistics of individual translation alternatives and statistics of query terms obtained using aggregated functions impact the ranking of documents.

5.4 Constraint Analysis on the Improved CLIR models

We first analyze the first modification of the PSQ method, obtained by changing the estimation of query term frequencies in Eq. 12. Since the estimation of term frequencies is only changed in this variant, we just investigate constraints regarding term frequencies.

Applying the assumptions of CL-C3 into Eq. 12, one has $c'(q, D_2) > c'(q, D_1)$. This is because the counts of the query term in the documents are calculated as follows:

$$c'(q, D_1) = (\sum_{t \in V_T} p(t|q) \times c(t, D_1)) \times \log(1 + \sigma), \tag{16}$$

$$c'(q, D_2) = (\sum_{t \in V_T} p(t|q) \times c(t, D_2)) \times \log(2 + \sigma), \tag{17}$$

where according to the assumptions of CL-C3, one has

$$\sum_{t \in V_T} p(t|q_a) \times c(t, D_1) = \sum_{t \in V_T} p(t|q_a) \times c(t, D_2). \tag{18}$$

Feeding the above query term frequencies into BM25 retrieval function, the new retrieval model thus prefers document D_2 over D_1, satisfying the CL-C3 constraint.

In Section 4.2, we showed that the original PSQ method satisfies the CL-C2 constraint when two query terms q_1 and q_2 have the same discrimination value. Based on this observation, we want to check the CL-C2 constraint on the new variant of the PSQ method. For this constraint to be satisfied, one should have the following inequality, obtained by applying the assumptions of the CL-C2 constraint into the PSQ+CL-C3 retrieval function.

$$\begin{aligned} &\frac{c(q_2, D_1) \times \log(2 + \sigma)}{k_1\big((1 - b) + b\frac{|D_1|}{\text{avdl}}\big) + c(q_2, D_1) \times \log(2 + \sigma)} \\ &\qquad - \frac{c(q_2, D_2) \times \log(1 + \sigma)}{k_1\big((1 - b) + b\frac{|D_2|}{\text{avdl}}\big) + c(q_2, D_2) \times \log(1 + \sigma)} \\ &\qquad < \frac{c(q_1, D_2) \times \log(1 + \sigma)}{k_1\big((1 - b) + b\frac{|D_2|}{\text{avdl}}\big) + c(q_1, D_2) \times \log(1 + \sigma)}. \end{aligned} \tag{19}$$

Note that the last component in the scoring function of BM25 model in Eq. 3, i.e., $((k_3 + 1) * c(q_i, \boldsymbol{q}))/(k_3 + c(q_i, \boldsymbol{q}))$, is equal to one because each query term has a frequency of one according to the assumptions of the constraint. Performing some simple algebraic operations, the inequality becomes:

$$\begin{aligned} &c(q_2, D_1) \times k_1\big((1 - b) + b\frac{|D_1|}{\text{avdl}}\big)^2 \times (\log(2 + \sigma) - \log(1 + \sigma)) \\ &\qquad < c(q_1, D_2) \times c(q_2, D_2) \times k_1\big((1 - b) + b\frac{|D_1|}{\text{avdl}}\big) \times \log^2(1 + \sigma) + \\ &\qquad\quad c(q_2, D_1) \times c(q_2, D_2) \times c(q_1, D_2) \times \log^2(1 + \sigma). \end{aligned} \tag{20}$$

Both query terms q_1 and q_2 have non-zero frequencies in document D_2, and one of them has a higher frequency than the other. Without loss of generality, we assume query term q_1 has the higher frequency in D_2, i.e., $c(q_2, D_2) \leq c(q_1, D_2)$. On the other hand, one has

$c(q_2, D_1) = c(q_2, D_2) + c(q_1, D_2)$ according to the assumptions of the CL-C2 constraint. Therefore, $c(q_2, D_1) < 2 \times c(q_1, D_2)$. Replacing $c(q_2, D_1)$ in Eq. 20 by its upper bound and performing some algebraic operations, one gets:

$$k_1\big((1-b) + b\frac{|D_1|}{\text{avdl}}\big)^2 < \frac{c(q_2, D_2) \times \log^2(1+\sigma) \times (k_1\big((1-b) + b\frac{|D_1|}{\text{avdl}}\big) + c(q_2, D_1))}{2 \times (\log(2+\sigma) - \log(1+\sigma))}, \tag{21}$$

where the right-hand side is increasing in the count $c(q_2, D_2)$. Therefore, the minimum of the right-hand side is obtained when $c(q_2, D_2)$ is set to its lowest value, which again makes the inequality harder to satisfy. The value of $c(q_2, D_2)$ is calculated using Eq. (1), and depends on translation frequencies in the document and their corresponding translation probabilities. The minimum frequency of a translation term in a document is 1. In addition, translation models are mostly generated by filtering translations with probabilities higher than a threshold, henceforth denoted by th. Therefore, the minimum frequency of a query term in a document is th. Replacing this minimum value in Eq. (21), the inequality becomes:

$$k_1\big((1-b) + b\frac{|D_1|}{\text{avdl}}\big)^2 < \frac{\mathsf{th} \times \log^2(1+\sigma) \times (k_1((1-b) + b\frac{|D_1|}{\text{avdl}}) + 2\mathsf{th})}{2 \times (\log(2+\sigma) - \log(1+\sigma))}. \tag{22}$$

In this final inequality, k_1 (usually between 1.0 and 2.0) and b (usually 0.75) are constants (Fang et al 2004). In addition, avdl is also a constant depending on the data collection. Appropriately setting the parameter σ according to these values, the above inequality will be satisfied for most documents in the collection. We later show in the experiments that, when $\sigma \geq 7$, more than 90% of documents in all collections satisfy the inequality (22). Note that this inequality is obtained by setting variables to their worst case values.

Constraints on term discrimination values. The PSQ+CL-C4 method does not generally satisfy the CL-C4 constraint, similar to the original PSQ method. However, when none of the translations occur in the intersection of two documents D_1 and D_2, i.e., $c(t_1, D) = c(t_2, D) = 0$ as shown in Fig. 3, the PSQ+CL-C4 satisfies the CL-C4 constraint, in contrast to the original PSQ method.

According to the assumptions of the CL-C4 constraint, $\text{dv}(t_1) > \text{dv}(t_2)$, therefore $\text{df}(t_1) < \text{df}(t_2)$. Substitution of this ordering into Eq. (14), given that document D_1 contains only translation t_1 and document D_2 has only translation t_2, yields $\text{df}(q, D_1) < \text{df}(q, D_2)$. The discrimination value of query term q in documents D_1 and D_2 using Eq. 15 become:

$$\text{dv}(q, D_1) = \ln(\frac{N+1}{\text{df}(q, D_1) + 0.5}) \times \frac{\ln(\frac{N+1}{\text{df}(q) - \text{df}(q, D_1) + 0.5})}{\ln(\frac{N+1}{\text{df}(q) - \text{df}(q, D_1) + 0.5}) + c}, \tag{23}$$

$$\text{dv}(q, D_2) = \ln(\frac{N+1}{\text{df}(q, D_2) + 0.5}) \times \frac{\ln(\frac{N+1}{\text{df}(q) - \text{df}(q, D_2) + 0.5})}{\ln(\frac{N+1}{\text{df}(q) - \text{df}(q, D_2) + 0.5}) + c}, \tag{24}$$

Table 2 Dataset properties.

Year	Data collection	Document language	Num of documents	Query language	Num of queries	Query Set Name
2002	Los Angeles Times 1994	English	113005	French	50	En-Fr’02
				Italian	50	En-It’02
				Spanish	50	En-Sp’02
	Le Monde 1994 French SDA 94	French	87191	English	50	Fr-En’02
	La Stampa 1994 Italian SDA 94	Italian	108578	English	50	It-En’02
2003	Los Angeles Times 1994 Glasgow Herald 1995	English	169477	French	60	En-Fr’03
				Spanish	60	En-Sp’03
2008 2009	Hamshahri	Persian	166774	English	100	Pr-En

Table 3 Performance of the PSQ, LM-based, and proposed HQM methods using 2-fold cross-validation. BM25 and LM-Dir are monolingual results as a baseline for the CLIR results. Statistical significant difference with the PSQ and the LM-based methods are denoted by $*$ and $\bullet$ symbols, respectively. We only show the significance tests for MAP measure.

	En-Fr’02		En-It’02		En-Sp’02		Fr-En’02	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
BM25	0.4063	0.3738	0.4063	0.3738	0.4063	0.3738	0.3407	0.332
LM-Dir	0.4176	0.3905	0.4176	0.3905	0.4176	0.3905	0.369	0.368
PSQ	0.3047	**0.3262**	0.2994	0.2690	0.2983	0.288	0.3169	0.286
LM-based	0.2971	0.3094	**0.3195**	**0.2977**	0.2823	0.288	**0.3306**	**0.3336**
HQM	**0.3243**$^{*\bullet}$	**0.3262**	0.3193*	0.2786	**0.3124**$^{*\bullet}$	**0.3167**	0.3271	0.332

	It-En’02		En-Fr’03		En-Sp’03		Pr-En	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
BM25	0.3289	0.3694	0.4564	0.3315	0.4564	0.3315	0.3676	0.5900
LM-Dir	0.3133	0.3776	0.4323	0.3111	0.4323	0.3111	0.3554	0.561
PSQ	0.2515$^{\bullet}$	0.2944	0.4014	0.3143	0.3687	**0.3111**	0.2344$^{\bullet}$	0.4
LM-based	0.2362	0.2803	0.4158	0.3339	0.3521	0.3074	0.1987	0.366
HQM	**0.2685**$^{\bullet}$	**0.3122**	**0.4466**$^{*\bullet}$	**0.3369**	**0.3989**$^{*\bullet}$	0.3093	**0.2422**$^{*\bullet}$	**0.405**

which show that $\text{dv}(q, D_1) > \text{dv}(q, D_2)$, since function $\frac{ln(x)}{ln(x)+c}$ is increasing for $x > 1$ and $\text{df}(q, D_1) < \text{df}(q, D_2)$. Having a higher discrimination value for q in document D_1, the PSQ+CL-C4 method prefers document D_1 in ranking, and thus satisfies the CL-C4 constraint.

Table 1 also summarizes the results of constraint analysis on the proposed model for CLIR.

6 Experiments

Datasets. Evaluations are carried out against test collections from ad-hoc cross-language track in CLEF-2002, CLEF-2003, CLEF-2008, and CLEF-2009 campaigns. We use English, French, Italian, and Persian collections with query sets in multiple languages for the conducted experiments, which represent different language pairs and different translation directions. Test collections and

Table 4 Performance of the proposed modifications on the PSQ method. The $*, \diamond$, and $\bullet$ symbols show statistical significant difference with the PSQ, PSQ+CL-C3, and PSQ+CL-C4 methods, respectively.

	En-Fr'02		En-It'02		En-Sp'02		Fr-En'02	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
PSQ	0.3084	**0.3357**	0.2853	0.2762	0.2837	0.2929	0.308	**0.318**
PSQ+CL-C3	0.31	0.3333	**0.2988**	**0.2786**	0.2968	0.2809	0.3108	**0.318**
% Impr	0.5%		4.7%		4.6%		0.9%	
PSQ+CL-C4	0.3161	0.3214	0.2869	0.2738	0.2937	0.2976	0.3202	0.308
% Impr	2.5%		0.5%		3.5%		4%	
HQM	**0.3301**	0.3262	0.2956	0.2714	**0.3054**	**0.3048**	**0.325**	0.316
% Impr	7%		3.6%		7.6%		5.5%	

	It-En'02		En-Fr'03		En-Sp'03		Pr-En	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
PSQ	0.2373	0.2837	0.3603	0.3093	0.3674	0.313	0.2255	0.365
PSQ+CL-C3	0.253	0.2966	0.3878^*	0.3209	$\mathbf{0.4004}^*$	**0.3148**	0.231^*	0.388
% Impr	6.6%		7.6%		9%		2.4%	
PSQ+CL-C4	0.2559^*	0.2923	0.3987^*	**0.3225**	0.3864	0.2981	0.2318^*	0.379
% Impr	7.8%		10.7%		5.2%		2.8%	
HQM	$\mathbf{0.266}^*$	**0.3082**	$\mathbf{0.4292}^{*\diamond\bullet}$	**0.3225**	0.3990	0.3056	$\mathbf{0.2427}^{*\diamond\bullet}$	**0.396**
% Impr	12.1%		19.1%		8.6%		7.6%	

their languages as well as query languages are shown in Table 2. In addition, the query set in the language of each collection is used to provide monolingual baseline for the performance of CLIR. For retrieval, we index the TEXT and TITLE fields of documents belonging to collections in European languages.

Preprocessing. Diacritic characters in European languages are mapped to the corresponding unmarked characters. Persian words are normalized by replacing all orthographic variations of letters by one form.[1] Stopwords are removed using stopword lists provided in *IR Multilingual Resources at UniNE*[2]. Persian stopword list provided by CLEF campaigns is used. We then use Snowball stemmers[3] for all European languages. Persian words are not stemmed, due to lack of a high quality stemmer.

Translation models. We build a word-to-word translation model for each European language pair using the *Europarl Corpus* (European Parliament Proceedings Parallel Corpus) version 7 (Tiedemann 2012). The translation model for English-Persian retrieval is built using TEP parallel corpus (Pilevar et al 2011). The GIZA++ toolkit (Och and Ney 2003) is used to obtain IBM model-1 translations. Before word alignment, the aforementioned preprocessing steps are done on both sides of each parallel corpus. After training using each parallel corpus, we impose a probability threshold of 0.1 to build a translation model (Nie 2010), where translation probabilities are re-normalized after filtering noise translations.

[1] Tool available at http://humanities.uva.nl/ dehghani/persian-linguistic-resources/.
[2] http://members.unine.ch/jacques.savoy/clef/
[3] http://snowball.tartarus.org/.

Parameter setting. We set the parameters of BM25 model and language modeling framework to the following default values, unless otherwise stated. For the BM25 model in the PSQ method, we use default parameter values as $k_1 = 1.2$, $b = 0.75$ and $k_3 = 7$, used in many studies such as (Wang and Oard 2012). The smoothing parameter of the LM-based model, μ, is set to the default value of 1000. In all experiments, the values of parameters σ and c (free parameters in Eqs. 12 and 15, respectively) are set using 2-fold cross-validation over the queries in each query set. K-fold cross-validation generally reduces the chance of overfitting, however parameters may still overfit. We varied the value of parameter σ between $\{1, 2, ..., 10\}$. The value of parameter c is selected from $\{0.5\} \cup \{1, 2, 3, \ldots, 20\}$.

For each experiment, we report Mean Average Precision (MAP) and Precision at top 10 documents (P@10). Two-tailed paired t-test is used to test whether the differences between performance of approaches are statistically significant.

7 Results and Discussion

In this section, we first compare the performance of PSQ and LM-based models through systematic experiments. Our goal is to see whether the results of empirical comparisons are consistent with those of analytical comparisons using the defined constraints. We then evaluate the performance of the proposed modifications to the PSQ method.

7.1 Performance Comparison between the PSQ and LM-Based Models

We first report the results of the PSQ and LM-based models on all datasets in Table 3. To provide a fair comparison, the parameters of the models are not set to the default values, but are set using the 2-fold cross-validation method. The parameters of the PSQ method, k_1 and b related to the BM25 method, are varied over $\{1, 1.2, 1.5, 1.75, 2\}$ and $\{0.25, 0.5, 0.75\}$ sets, respectively. The parameter of the LM-based retrieval is due to the smoothing method used in the estimation of document language models, for which we used the Dirichlet-prior smoothing method. The parameter μ of this smoothing method is set by examining the values in $\{100, 200, 500, 1000, 2000\}$.

The results of Table 3 show that neither of the PSQ and LM-based models outperforms the other in all datasets. This finding is compatible with the results of analytical comparison of the two models in Table 1, where each of the two models satisfies a subset of the defined constraints on CLIR models.

The alteration of the better retrieval model across different datasets shows that the relative importance of different constraints is not fixed, and indeed depends on the characteristics of the datasets and queries (e.g., language of queries and documents). However, in most cases, the differences between the performance of these two models are not statistically significant. Although performance differences between the two models are not statistically significant in most cases, these differences are substantial for some datasets. For the Hamshahri dataset, the PSQ method greatly outperforms the LM-based model. We hypothesize that the reason of this remarkable difference between the performance of the two models stems from the lower quality of the translations between English-Persian languages. Overall, the lower quality of translations between English-Persian is evident from the lower percentage of the CLIR performance over the monolingual baseline on the Hamshahri dataset compared to that on other datasets. The result for the Hamshahri dataset shows that in the case of low-quality translations, the performance of LM-based model can be hurt more. This is mainly due to the fact that the LM-based model does not satisfy the CL-C2 constraint about covering translations of more distinct query terms (see Table 1), and therefore this constraint can have a detrimental effect on the performance of the LM-based method when translations have low quality.

7.2 Performance of the Proposed Modifications

We modify the PSQ method in two steps, and therefore, we first investigate the impact of each improvement separately, and then provide the performance of the final retrieval model by considering both updates.

In the first step, we improve the estimation of term frequency in the PSQ method so that it satisfies the CL-C3 constraint regarding the frequency of different translations of a query term. In Table 4, the PSQ method is compared with the provided variant. These results are obtained by using the default values for the parameters of the BM25 retrieval model. The modified PSQ method consistently outperforms the original one across all datasets, and the improvements are statistically significant for 3 datasets.

Next, we propose the PSQ+CL-C4 variant that satisfies the CL-C4 constraint more than the original PSQ method. In Table 4, the PSQ method is compared with the proposed PSQ+CL-C4 variant. These results are also obtained by using default values for the parameters of the BM25 retrieval model, which show that PSQ+CL-C4 variant outperforms the PSQ method in all datasets in MAP measure. Improvements of the PSQ+CL-C3 and PSQ+CL-C4 variants over the PSQ method are very close in most datasets (see Table 4), which show that the two corresponding constraints are important and have an approximately equal impact on the performance with the proposed modifications. For En-Fr'03 and Pr-En datasets, improvements of both PSQ+CL-C3 and PSQ+CL-C4 variants are statistically significant over the PSQ method, while for some others, improvement by considering only one of the constraints is statistically significant. Finally, satisfaction of the CL-C3 or CL-C4 constraints by a CLIR model improves the retrieval performance.

We also propose the HQM method by considering both modifications. The results obtained by this method are reported in Table 4, which show that the HQM method outperforms the PSQ method in all datasets, and improves

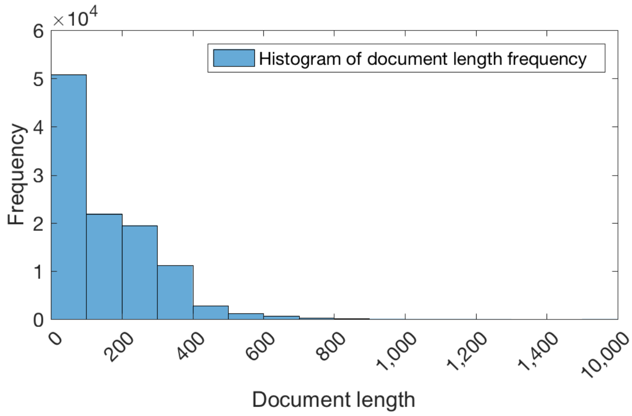


Fig. 4 Distribution of document length in La Stampa 1994 and Italian SDA 94 collections

the CLIR performance by 3.6% to 19.1% (which is substantial and statistically significant). Comparing the results of the HQM method with those of the individual modifications reveals that the HQM method outperforms the PSQ+CL-C4 method in all datasets, and outperforms the PSQ+CL-C3 method in all datasets except En-It'02 and En-Sp'03 datasets. For these two datasets, PSQ+CL-C3 shows the highest CLIR performance. Improvements of the HQM method in En-Fr'03 and Pr-En datasets are statistically significant over all variants, the original PSQ, PSQ+CL-C3, and PSQ+CL-C4 methods, although the improvements of both PSQ+CL-C3 and PSQ+CL-C4 over the PSQ method are statistically significant.

Finally, the HQM method is compared against the existing PSQ and LM-based methods in Table 3 using 2-fold cross-validation for selecting the parameter values of each method. First, the HQM method outperforms the PSQ method in all datasets, where the improvements are statistically significant in most cases. In addition, the HQM method shows statistically significant improvements over the LM-based method in most datasets, even for the cases that the performance difference between the PSQ and LM-based methods are not statistically significant. Finally, the HQM method achieves the highest CLIR performance in comparison with the PSQ and LM-based methods in most datasets. For the Fr-En'02 dataset, none of the methods shows statistically significant differences with others.

7.3 Empirical Evaluation of HQM with respect to the CL-C2 Constraint

We proved in Section 5.4 that PSQ+CL-C3 and thus the HQM method satisfy the CL-C2 constraint when Eq. (22) is satisfied. This inequality provides the minimal cases that the proposed modification satisfies the CL-C2 constraint, since Eq. (22) is obtained by considering the worst-case values for some parameters. The obtained inequality basically puts a constraint on document

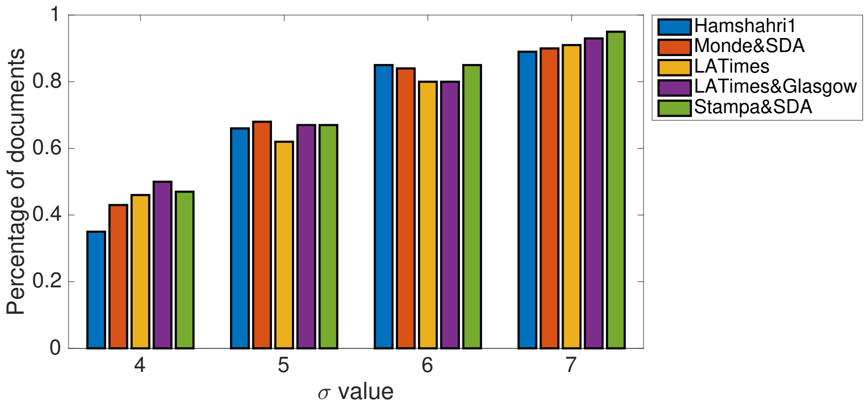


Fig. 5 Percentage of documents in each collection that satisfy the CL-C2 constraint on the HQM method, calculated for different values of σ in Eq. 22

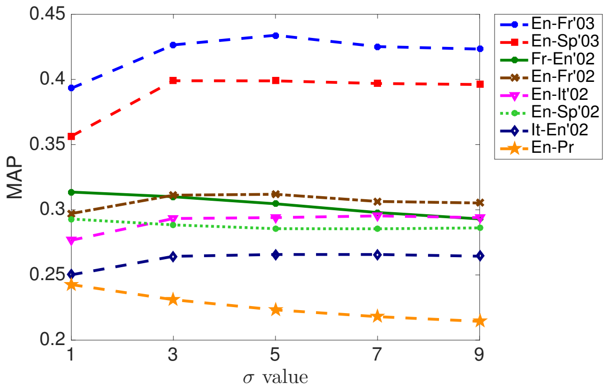


Fig. 6 Sensitivity of the performance of the HQM method to the parameter σ.

lengths. Roughly speaking, for documents with short lengths in a collection, the CL-C2 constraint will be satisfied with a small value for parameter σ. However, for long documents, we need to set parameter σ to a higher value to satisfy the constraint. Fig. 4 shows the frequency of document lengths in the collection of La Stampa 1994 and Italian SDA 94. The distribution shows that 47% of documents in this collection have length smaller than 100, thus the CL-C2 constraint will be satisfied for these documents with $\sigma = 4$. Therefore, with a small value for parameter σ, approximately 50% of the documents satisfy the CL-C2 constraints.

We evaluate inequality (22) with respect to the documents in our test collections to investigate how many documents in each collection satisfy the inequality (22). For a detailed analysis, we set the other parameters as $\mathsf{th} = 0.1$ (translation probability), $b = 0.75$ and $k_1 = 1$ (parameters of the BM25

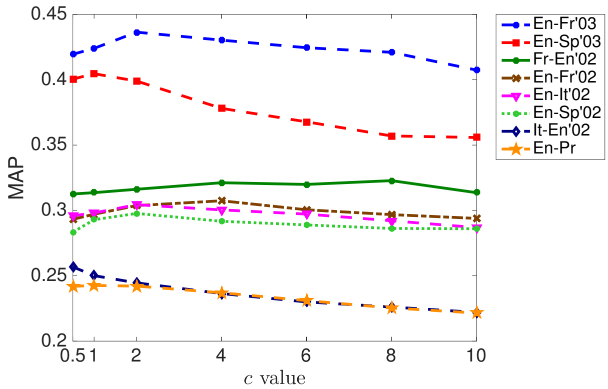


Fig. 7 Sensitivity of the proposed HQM method to the parameter c.

method). However, we change the value of parameter σ, the free parameter introduced in the proposed modification, to study how it impacts the satisfaction of the CL-C2 constraint by the HQM method. Fig. 5 shows the percentage of documents in each collection that satisfy the inequality (22) for different values of parameter σ. The figure shows that the percentage of documents that satisfy the inequality increases with the increase in the value of parameter σ, and when $\sigma = 7$, almost 90% of documents in all collections satisfy the inequality (22). Therefore, the HQM method satisfies the CL-C2 constraint for most cases. Finally, if the value of parameter σ is changed over $\{1, 2, \ldots, 10\}$, one can find the suitable parameter value so that the HQM method satisfies the CL-C2 constraint.

We also study the impact of parameter σ on the MAP performance of the HQM method in Fig. 6. Logically, the retrieval performance should increase with the satisfaction percentage of documents in a collection. The performance of 5 out of 8 datasets in Fig. 6 increases with the increase in the value of parameter σ, which conforms with the percentage depicted in Fig. 5 for these datasets. The best results for remaining datasets are obtained when $\sigma = 1$. The reason can be that the percentages depicted in Fig. 5 are based on inequality (22) obtained by considering the worst-case values for some parameters, which can be different in real collections. The other reason for this observation could be about the number of documents considered in these two diagrams. For Fig. 5, all documents in collections are considered for evaluation, while for measuring MAP performance in Fig. 6, only the top 1,000 documents with respect to each query are evaluated. The majority of these top 1,000 documents satisfy inequality (22) when $\sigma = 1$ since these documents are not very long, and thus increasing the value of parameter σ can hurt the performance of retrieval.

7.4 Sensitivity of the HQM Method with respect to Parameters

In this section, we study the sensitivity of the HQM method with respect to parameter c used in the estimation of term discrimination values in Eq. (15). Fig. 7 shows the CLIR performance for different values of parameter c across all datasets. The figure shows that the highest MAP performance for most datasets is achieved when c has a value between 1 and 4. In addition, the suitable value for parameter c appears to be more dependent on collection properties than query terms. Based on the figure, the curves for En-Fr'03 and En-Sp'03 datasets, which share the same collection of documents, have similar shapes. Similarly, the curves for En-Fr'02, En-It'02, and En-Sp'02 datasets have very similar shapes. Dependence of the value of parameter c, which is related to the estimation of term discrimination values, on collection properties is very reasonable, since this factor is used to discriminate terms of a collection for better retrieval. Finally, the value of parameter c should not be too small or too large, otherwise translations of query terms not occurred in documents or query term discrimination values, respectively, are ignored in document ranking.

8 Conclusion and Future Work

In this paper, we provide formal representations of the impacts of translation knowledge on retrieval heuristics and, consequently, on the ranking of documents in CLIR. The defined constraints concern the impacts of term frequency and term discrimination values, which are estimated using translation knowledge in the CLIR setting. Regarding each heuristic, constraints are defined on two levels of statistics. In particular, constraints on term frequency considered frequency of translations in documents, and frequency of a query term in documents obtained using an aggregated function on frequency of its translations. Similarly, these two statistics are considered in constraints about query term discrimination values.

The defined constraints are checked against the existing methods for CLIR, and we showed that none of them fully satisfies all the constraints. Therefore, the proposed constraints pave the path for improving the existing methods to achieve a higher CLIR performance. In this line, we proposed hierarchical query modeling based on the probabilistic structured query method for CLIR, that satisfies more constraints. Empirical evaluation of the existing and proposed methods demonstrate that the proposed modifications achieve a higher CLIR performance over all test datasets. This shows the importance of satisfying the defined constraints on CLIR methods.

Our proposed constraints and their empirical effects stimulate further research. First, the proposed constraints are focused on basic retrieval heuristics and there are several aspects affecting document ranking that can also be formalized through constraints. This can help to further improve the existing methods for CLIR. Another promising direction is to have a CLIR method that satisfies all constraints. The proposed modifications can also be improved to get higher CLIR performance. Finally, multilingual information retrieval is even more challenging than cross-language information retrieval, and research about how different factors including adopting different translation knowledge should impact document ranking is very interesting.

9 Acknowledgements

We are grateful to the anonymous reviewers for their constructive comments. This research was supported in part by a grant from the Institute for Research in Fundamental Sciences (No. CS1399-4-286), and in part by the Center for Intelligent Information retrieval.

References

Amigó E, Gonzalo J, Verdejo F (2013) A general evaluation measure for document organization tasks. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '13, pp 643–652, DOI 10.1145/2484028.2484081, URL http://doi.acm.org/10.1145/2484028.2484081

Ariannezhad M, Montazeralghaem A, Zamani H, Shakery A (2017a) Improving retrieval performance for verbose queries via axiomatic analysis of term discrimination heuristic. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '17, pp 1201–1204, DOI 10.1145/3077136.3080761, URL http://doi.acm.org/10.1145/3077136.3080761

Ariannezhad M, Montazeralghaem A, Zamani H, Shakery A (2017b) Iterative estimation of document relevance score for pseudo-relevance feedback. In: Proceedings of the IR research, 39th European conference on Advances in information retrieval, Springer-Verlag, ECIR'17, pp 676–683

Berger A, Lafferty J (1999) Information retrieval as statistical translation. In: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '99, pp 222–229, DOI 10.1145/312624.312681, URL http://doi.acm.org/10.1145/312624.312681

Buckley C, Mitra M, Walz J, Cardie C (2000) Using clustering and superconcepts within smart: Trec 6. Inf Process Manage 36(1):109–131, DOI 10.1016/S0306-4573(99)00047-3, URL http://dx.doi.org/10.1016/S0306-4573(99)00047-3

Busin L, Mizzaro S (2013) Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In: Proceedings of the 2013 Conference on the Theory of Information Retrieval, ACM, New York, NY, USA, ICTIR '13, pp 8:22–8:29, DOI 10.1145/2499178.2499182, URL http://doi.acm.org/10.1145/2499178.2499182

Clinchant S, Gaussier E (2011) Is document frequency important for prf? In: Proceedings of the Third International Conference on Advances in Information Retrieval Theory, Springer-Verlag, Berlin, Heidelberg, ICTIR'11, pp 89–100, URL http://dl.acm.org/citation.cfm?id=2040317.2040331

Clinchant S, Gaussier E (2013) A theoretical analysis of pseudo-relevance feedback models. In: Proceedings of the 2013 Conference on the Theory of Information Retrieval, ACM, New York, NY, USA, ICTIR '13, pp 6:6–6:13, DOI 10.1145/2499178.2499179, URL http://doi.acm.org/10.1145/2499178.2499179

Darwish K, Oard DW (2003) Probabilistic structured query methods. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, ACM, New York, NY, USA, SIGIR '03, pp 338–344, URL http://doi.acm.org/10.1145/860435.860497

Fang H (2008) A re-examination of query expansion using lexical resources. In: Proceedings of ACL-08: HLT, Association for Computational Linguistics, Columbus, Ohio, pp 139–147, URL http://www.aclweb.org/anthology/P/P08/P08-1017

Fang H, Zhai C (2005) An exploration of axiomatic approaches to information retrieval. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '05, pp 480–487, DOI 10.1145/1076034.1076116, URL http://doi.acm.org/10.1145/1076034.1076116

Fang H, Zhai C (2006) Semantic term matching in axiomatic approaches to information retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '06, pp 115–122, DOI 10.1145/1148170.1148193, URL http://doi.acm.org/10.1145/1148170.1148193

Fang H, Tao T, Zhai C (2004) A formal study of information retrieval heuristics. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '04, pp 49–56, URL http://doi.acm.org/10.1145/1008992.1009004

Fang H, Tao T, Zhai C (2011) Diagnostic evaluation of information retrieval models. ACM Trans Inf Syst 29(2):7:1–7:42, DOI 10.1145/1961209.1961210, URL http://doi.acm.org/10.1145/1961209.1961210

Gerani S, Zhai C, Crestani F (2012) Score transformation in linear combination for multi-criteria relevance ranking. In: Proceedings of the 34th European Conference on Advances in Information Retrieval, Springer-Verlag, Berlin, Heidelberg, ECIR'12, pp 256–267

He D, Oard DW, Wang J, Luo J, Demner-Fushman D, Darwish K, Resnik P, Khudanpur S, Nossal M, Subotin M, Leuski A (2003) Making miracles: Interactive translingual search for cebuano and hindi. ACM Trans Asian Lang Inf Process 2(3):219–244, DOI 10.1145/979872.979876, URL http://doi.acm.org/10.1145/979872.979876

Karimzadehgan M, Zhai C (2012) Axiomatic analysis of translation language model for information retrieval. In: Proceedings of the 34th European Conference on Advances in Information Retrieval, Springer-Verlag, Berlin, Heidelberg, ECIR'12, pp 268–280

Kern R, Juffinger A, Granitzer M (2009) Evaluation of axiomatic approaches to crosslanguage retrieval. In: Proceedings of the 10th Cross-language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments, Springer-Verlag, Berlin, Heidelberg, CLEF'09, pp 142–149, URL http://dl.acm.org/citation.cfm?id=1887364.1887386

Kraaij W, de Jong F (2004) Transitive probabilistic CLIR models. In: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, Paris, France, France, RIAO '04, pp 69–81, URL http://dl.acm.org/citation.cfm?id=2816272.2816280

Kraaij W, Nie JY, Simard M (2003) Embedding web-based statistical translation models in cross-language information retrieval. Comput Linguist 29(3):381–419, DOI 10.1162/089120103322711587, URL http://dx.doi.org/10.1162/089120103322711587

Lavrenko V, Choquette M, Croft WB (2002) Cross-lingual relevance models. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '02, pp 175–182, DOI 10.1145/564376.564408, URL http://doi.acm.org/10.1145/564376.564408

Li B, Gaussier E (2012) An information-based cross-language information retrieval model. In: Proceedings of the 34th European Conference on Advances in Information Retrieval, Springer-Verlag, Berlin, Heidelberg, ECIR'12, pp 281–292

Li B, Gaussier E, Yang D (2018) The dilution/concentration conditions for cross-language information retrieval models. Information Processing & Management 54(2):291–302

Lv Y (2015) A study of query length heuristics in information retrieval. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM '15, pp 1747–1750, DOI 10.1145/2806416.2806592, URL http://doi.acm.org/10.1145/2806416.2806592

Lv Y, Zhai C (2011) Lower-bounding term frequency normalization. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM '11, pp 7–16, DOI 10.1145/2063576.2063584, URL http://doi.acm.org/10.1145/2063576.2063584

Mcnamee P, Mayfield J (2004) Character n-gram tokenization for european language text retrieval. Inf Retr 7(1-2):73–97, DOI 10.1023/B:INRT.0000009441.78971.be, URL http://dx.doi.org/10.1023/B:INRT.0000009441.78971.be

Montazeralghaem A, Zamani H, Shakery A (2016) Axiomatic analysis for improving the log-logistic feedback model. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '16, pp 765–768, DOI 10.1145/2911451.2914768, URL http://doi.acm.org/10.1145/2911451.2914768

Montazeralghaem A, Zamani H, Shakery A (2017) Term proximity constraints for pseudo-relevance feedback. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '17, pp 1085–1088, DOI 10.1145/3077136.3080728, URL http://doi.acm.org/10.1145/3077136.3080728

Na SH, Kang IS, Lee JH (2008) Improving term frequency normalization for multi-topical documents and application to language modeling approaches. In: Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval, Springer-Verlag, Berlin, Heidelberg, ECIR'08, pp 382–393, URL http://dl.acm.org/citation.cfm?id=1793274.1793321

Nie JY (2010) Cross-Language Information Retrieval. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers

Oard DW, Wang J (2001) NTCIR-2 ECIR experiments at maryland: Comparing pirkola's structured queries and balanced translation. In: Proceedings of the Third Second Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, NTCIR-2

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51, DOI 10.1162/089120103321337421, URL http://dx.doi.org/10.1162/089120103321337421

Pal D, Mitra M, Bhattacharya S (2015) Improving pseudo relevance feedback in the divergence from randomness model. In: Proceedings of the 2015 International Conference on The Theory of Information Retrieval, ACM, New York, NY, USA, ICTIR '15, pp 325–328, DOI 10.1145/2808194.2809494, URL http://doi.acm.org/10.1145/2808194.2809494

Pilevar MT, Faili H, Pilevar AH (2011) TEP: Tehran English-Persian parallel corpus. In: Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part II, Springer-Verlag, Berlin, Heidelberg, CICLing'11, pp 68–79, URL http://dl.acm.org/citation.cfm?id=1964750.1964757

Rahimi R, Shakery A, King I (2014) Axiomatic analysis of cross-language information retrieval. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM '14, pp 1875–1878, DOI 10.1145/2661829.2661915, URL http://doi.acm.org/10.1145/2661829.2661915

Savoy J (2005) Comparative study of monolingual and multilingual search models for use with asian languages. ACM Trans Asian Lang Inf Process 4(2):163–189, DOI 10.1145/1105696.1105701, URL http://doi.acm.org/10.1145/1105696.1105701

Sorg P, Cimiano P (2012) Exploiting wikipedia for cross-lingual and multilingual information retrieval. Data Knowl Eng 74:26–45, DOI 10.1016/j.datak.2012.02.003, URL http://dx.doi.org/10.1016/j.datak.2012.02.003

Tao T, Zhai C (2007) An exploration of proximity measures in information retrieval. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '07, pp 295–302, DOI 10.1145/1277741.1277794, URL http://doi.acm.org/10.1145/1277741.1277794

Tiedemann J (2012) Parallel data, tools and interfaces in opus. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey

Vulić I, Moens MF (2015) Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: Proceedings of the 38th International ACM SIGIR

Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '15, pp 363–372, DOI 10.1145/2766462.2767752, URL http://doi.acm.org/10.1145/2766462.2767752

Wang J, Oard DW (2012) Matching meaning for cross-language information retrieval. Inf Process Manage 48(4):631–653, DOI 10.1016/j.ipm.2011.09.003, URL http://dx.doi.org/10.1016/j.ipm.2011.09.003

Wu H, Fang H (2012) Relation based term weighting regularization. In: Proceedings of the 34th European Conference on Advances in Information Retrieval, Springer-Verlag, Berlin, Heidelberg, ECIR'12, pp 109–120, DOI 10.1007/978-3-642-28997-2_10, URL http://dx.doi.org/10.1007/978-3-642-28997-2_10

Xu J, Weischedel R, Nguyen C (2001) Evaluating a probabilistic model for cross-lingual information retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, SIGIR '01, pp 105–110, DOI 10.1145/383952.383968, URL http://doi.acm.org/10.1145/383952.383968

Zheng W, Fang H (2010) Query aspect based term weighting regularization in information retrieval. In: Proceedings of the 32nd European Conference on Advances in Information Retrieval, Springer-Verlag, Berlin, Heidelberg, ECIR'10, pp 344–356