Term Discrimination Value for Cross-Language Information Retrieval

Ali Montazeralghaem, Razieh Rahimi, and James Allan
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
Amherst, MA 01003
{montazer,rahimi,allan}@cs.umass.edu

ABSTRACT

Term discrimination value is among the three basic heuristics exploited, directly or indirectly, in almost all ranking models for adhoc Information Retrieval (IR). Query term discrimination in monolingual IR is usually estimated based on document or collection frequency of terms. In query translation approach for CLIR, discrimination value of a query term needs to be estimated based on document or collection frequencies of its translations, which is more challenging. We show that the existing estimation models do not correctly estimate and adequately reflect the difference between discrimination power of query terms, which hurts the retrieval performance. We then propose a new model to estimate discrimination values of query terms for CLIR and empirically demonstrate its impact in improving the CLIR performance.

KEYWORDS

Cross-language information retrieval, term discrimination value, probabilistic structured query

ACM Reference Format:
Ali Montazeralghaem, Razieh Rahimi, and James Allan. 2019. Term Discrimination Value for Cross-Language Information Retrieval. In *The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19), October 2–5, 2019, Santa Clara, CA, USA.* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3341981.3344252

1 INTRODUCTION

Cross-Language Information Retrieval (CLIR) is the task of retrieving documents with respect to queries in a language different than the language of documents. The general approach for crossing the language barrier between queries and the documents, is to use some sort of translation. Following the dominant approach for translation-based CLIR, we focus on the query translation approach [15]. Machine-readable bilingual dictionaries do not provide sufficient coverage for CLIR due to out of vocabulary words and neologisms. To compensate this deficiency, CLIR models tend to use statistical translation models learned from aligned bilingual corpora, to achieve acceptable performance. These models are referred to as corpus-based CLIR models.

Two widely-used basic methods for corpus-based CLIR are: (1) The probabilistic structured query (PSQ) method [5], and (2) cross-language information retrieval based on the language modeling framework [11, 12, 24] (henceforth referred to as the LM-based method). While two approaches have comparable results, and there is no constant winner over all test datasets, the PSQ performs better than the LM-based models in more cases [19]. The PSQ method provides two statistical estimates for each query term; term frequency in a document and document frequency. These estimates can then be adopted in any monolingual retrieval model based on these statistics of query terms, such as BM25 model [20], to rank documents in CLIR.

The PSQ method has shown promising results, yet we show that there is still one issue in the estimation of document frequency of query terms that limits the performance of the PSQ method. More specifically, document frequency of a query term is estimated based on the document frequency of its translations, where each translation is weighted by the translation probability obtained from a translation resource. The estimated document frequency is then used to compute the discrimination value of the query term. We show that between two query terms, the one that has a translation with lower document frequency and higher translation probability may incorrectly get a lower discrimination value. We propose a modification to the PSQ to solve this issue. Our experimental results on multiple standard datasets show that our proposed modification on the PSQ method significantly improves the performance of CLIR.

2 RELATED WORK

The task of CLIR is to score documents with respect to a query in a language different 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 can be performed without any direct translation [2, 4, 8, 14, 21]. 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 representation 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 [10, 22, 23]. The low-dimensional vectors for the CLIR task, proposed by Vulić and Moens [23], can be considered as an intermediate language space for queries and documents, because word

ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6881-0/19/10…$15.00
https://doi.org/10.1145/3341981.3344252

Table 1: An example query and estimated document frequencies using the PSQ method [5]

Query term	Translations		PSQ-df
	Trans. prob.	Mono-df	
q_1	0.8	100	200,080
	0.2	1,000,000	
q_2	0.8	10,000	198,000
	0.2	950,000	

embedding is used for representation of words in documents and queries which are in two languages. Each strategy has its own advantages and limitations.

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 [3], to CLIR. The cross-lingual models proposed in [11, 12, 24] belong to the first category. More specifically, Xu et al. [24] used a general collection in the query language for smoothing the new estimated language models for documents, while Kraaij et al. [11] 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, probabilistic structured query model proposed in [5] 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 [16]. They empirically compare the performance of the PSQ method with balanced translation for English-Chinese information retrieval, and show that the PSQ method outperforms balanced translation.

Li and Gaussier [13] 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. Learning-based models for CLIR have shown promising results [1, 7, 18]. They all use term discrimination value estimated by the PSQ method as a feature to represent queries. Replacing this IDF feature in these learning-based models with a better estimate will lead to better CLIR performance.

3 THE PROPOSED PSQ++ METHOD

We first describe the Probabilistic Structured Query method which is the basis of our proposed model, and then describe the proposed modification to the original PSQ method.

PSQ method. *Probabilistic Structured Queries* [5] is among the representative basic ranking models for cross-language information retrieval. Given a probabilistic translation model, frequency of a query word in a document written in another language is estimated as follows:

$$\mathrm{c}(q_i, d) = \sum_{w \in V_t} p(w|q_i)\mathrm{c}(w, d), \tag{1}$$

where w is a term in the vocabulary of document's language V_t, and $p(w|q_i)$ is the probability of translating word q_i into word w. Similarly, document frequency of a query term is estimated using document frequency of translations as

$$\mathrm{df}(q_i) = \sum_{w \in V_t} p(w|q_i)\mathrm{df}(w_t). \tag{2}$$

These frequency estimates are then used in a monolingual retrieval model to score documents, where BM25 has been mainly used. The BM25 model uses *inverse document frequency* (IDF) to discriminate query terms. There are multiple ways to calculate IDF from document frequency of terms, for which we use the following in our study [6]:

$$\mathrm{idf}(w) = \log \frac{N+1}{\mathrm{df}(w)}, \tag{3}$$

where N is the total number of documents in the corpus.

PSQ++ method. Before we describe the proposed improvement on the PSQ method, let's first consider an example query shown in Table 1 and see how its terms are distinguished using the PSQ method. The query has two query terms q_1 and q_2, where each query term has two translations with probabilities and document frequencies mentioned in the table. Both query terms q_1 and q_2 translate with probability 0.2 to a term that occurs very frequently in the corpus, i.e., in 1,000,000 and 950,000 documents, respectively. These translations, even if correct translations, cannot discriminate relevant and non-relevant documents because of their high document frequencies. On the other hand, query term q_1 translates with probability 0.8 to a term with document frequency of 100, while translation of q_2 with the same probability occurs in 10,000 documents. Therefore, query term q_1 should have a higher discrimination weight (IDF) than q_2 in ranking the documents. However, as shown in the table, the IDF values computed by the PSQ method have the reverse order.

Generally, a query term whose translations with high translation probabilities have low document frequencies is expected to have a high IDF value, which means to have a low aggregated document frequency (according to Eq. 3). Therefore, low document frequency is expected for a query term with low document frequency translations that have high translation probabilities. This implies that document frequency of a query term should have a negative correlation with translation probabilities, but a positive correlation with document frequencies of translations. However, as in Eq. 2, the document frequency of translations and their probabilities are multiplied together, thus they both have positive correlations with estimated document frequency for a query term, which can cause incorrect IDF weights in some cases such as the example in Table 1.

Another potential flaw of the document frequency estimation in the PSQ method is that translation probabilities with values in (0, 1] range can be dominated by large values of document frequencies in big corpora, allowing highly frequent translations determine the discrimination value of the query term, even when they have low translation probabilities.

To address the weaknesses of the PSQ method, we propose to estimate discrimination values of a query term as weighted combination of discrimination values of its translations, as follows:

$$\mathrm{idf}(q_i) = \sum_{w_t \in V_t} p(w_t|q_i)\mathrm{idf}(w_t), \tag{4}$$

Table 2: Datasets.

Data collection	Document language	Query language	Experiment name	Num of queries
LA Times 1994	English	French	qFr-docEn	50
		Italian	qIt-docEn	50
Le Monde 1994 French SDA 94	French	English	qEn-docFr	50
La Stampa 1994 Italian SDA 94	Italian	English	qEn-docIt	50

Table 3: MAP and precision performance of monolingual information retrieval using BM25.

qEn-docEn		qFr-docFr		qIt-docIt	
MAP	P@10	MAP	P@10	MAP	P@10
0.4063	0.3738	0.3407	0.3320	0.3289	0.3694

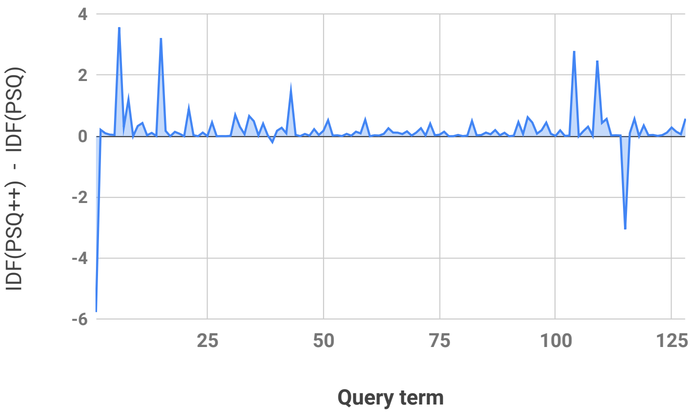

Figure 1: IDF values of each query term.

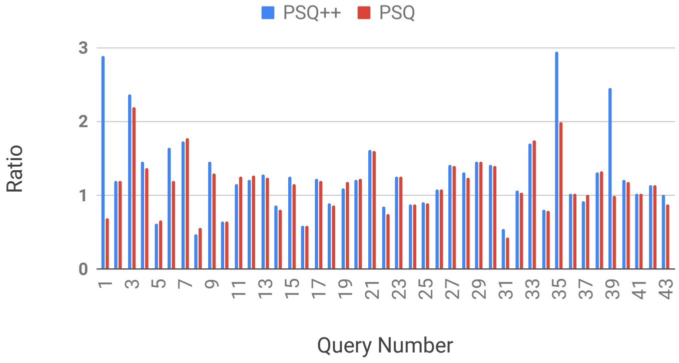

Figure 2: Ratio of IDF of query terms for eah query.

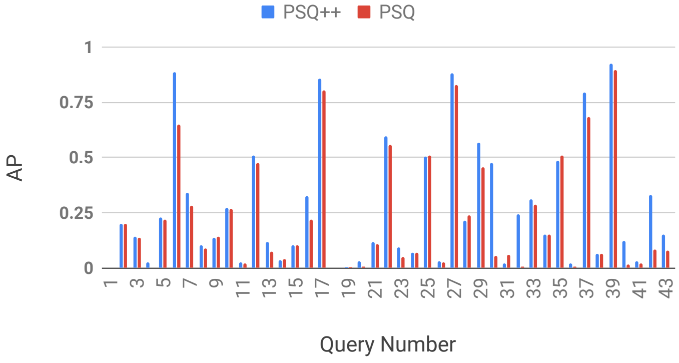

Figure 3: Average precision of each query.

where $\mathrm{idf}(w_t)$ is estimated using Eq. (3) using document frequency of translation w_t, which can be directly computed using the collection of documents. This estimate of IDF values for query terms in the cross-language setting does not have the issues of the estimate in Eq. 2, and can be used in retrieval models to weight query terms.

4 EXPERIMENTS

Datasets and experimental setup. We perform experiments on test collections from ad-hoc cross-language track in CLEF-2002 campaigns. We use English, French, and Italian collections to cover CLIR for different language pairs and different translation directions. Table 2 shows the datasets used in our experiments. Experiments are done using the Lemur toolkit[1].

Diacritic characters are mapped to the corresponding unmarked characters. Stopwords are removed using stopword lists provided in *IR Multilingual Resources at UniNE*[2]. Next, words of all languages are stemmed using Snowball stemmers[3]. The TEXT and TITLE fields of documents in test collections are then indexed for retrieval. We use a word-to-word translation model (IBM model 1) for each language pair learned on the Europarl corpus [9] by GIZA++ toolkit [17]. Both sides of each parallel corpus are preprocessed before word alignment. We use the top 3 translations for each word in our experiments, and translation probabilities are linearly normalized.

For each experiment, we report Mean Average Precision (MAP) and Precision at top 10 documents (P@10). Two-tailed paired t-test at a 95% confidence level is performed to test whether the differences between MAP performance of PSQ and PSQ++ are statistically significant. We also show the results of LM-based method [12] in this table as another baseline. Table 3 shows the performance of monolingual retrieval on the document collection of datasets as a baseline for cross-language information retrieval.

Table 4 shows the performance of CLIR when discrimination values of query terms are estimated using the PSQ method and the proposed PSQ++ method. As the results show, the proposed estimation improves the retrieval performance across all datasets, and the improvements are statistically significant.

For "qEn-docIt" dataset, we show more detailed results. First, Figure 1 show that although the two estimates of term discrimination values in PSQ and proposed PSQ++ methods seem to be similar, they are considerably different for some query terms such as the

[1] https://lemurproject.org/lemur.php
[2] http://members.unine.ch/jacques.savoy/clef/
[3] http://snowball.tartarus.org/.

Table 4: MAP and precision performance of the CLIR methods.

Method	qFr-docEn		qIt-docEn		qEn-docFr		qEn-docIt	
	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10
LM-based	0.3055	0.3146	0.3110	0.2805	0.3129	**0.3184**	0.2281	0.2750
PSQ	0.2732	0.3122	0.2785	0.2537	0.2573	0.2857	0.2077	0.2562
PSQ++	**0.3242**	**0.3488**	**0.3167**	**0.2951**	**0.3162**	**0.3184**	**0.2502**	**0.2812**

first query term. Although the absolute values of IDF of query terms impact the retrieval performance, we hypothesize that the ratio of IDF of different query terms of a query is more important for retrieval effectiveness. To study how PSQ and PSQ++ impact IDF ratio between query terms of a query term, we average the IDF ratio of consecutive pairs of query terms. More specifically, given a query $q = \{q_1, q_2, \ldots, q_n\}$, we compute the following value for the query.

$$\frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\mathrm{idf}(q_{i+1})}{\mathrm{idf}(q_i)}. \tag{5}$$

The $\mathrm{idf}(q_i)$ in this equation is estimated using PSQ or PSQ++ method. Queries that have a query term with estimated IDF value of zero are removed from consideration. This happens when none of the translations of a query term occurs in the corpus. Figure 2 shows the averaged ratio values for each query. As shown in the figure, the ratio of query terms estimated by the two methods are quite different. We also observe that the ranking of query terms by their IDF values estimated by PSQ and PSQ++ are different for some queries. To investigate how difference in IDF ratio of query terms impacts retrieval performance, we provide the average precision of queries in the dataset in Figure 3. For query 6 in the figure, we show that the PSQ++ method outperforms the PSQ. Based on Figure 2, one can observe that IDF ratios of query terms in query 6 by the PSQ and PSQ++ methods are considerably different. This observation conforms our hypothesis that IDF ratio of query terms impact the retrieval performance. In addition, the higher CLIR performance using the PSQ++ method shows that the PSQ++ can better reflect the difference between discrimination power of query terms.

5 CONCLUSION AND FUTURE WORK

We show that the estimation of document frequency for query terms in the *probabilistic structured query* has an issue that hinders the performance of CLIR. The issue is that the estimated document frequency for a query term has a positive correlation with the translation probability, while high translation probability indicate a high quality translation that its discrimination power should not decrease by its translation probability. We proposed a modification to the PSQ method to better estimate the IDF values of query terms and show the empirical impacts of the proposed improvement on the performance of CLIR. One interesting direction for future research is to compare the two methods based on axiomatic analysis framework, where constraints that a method for IDF estimation in CLIR models should statisfy in order to provide reasonable rankings of documents are formulated.

Acknowledgements. This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-1617408. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Hosein Azarbonyad, Azadeh Shakery, and Heshaam Faili. 2012. Using Learning to Rank Approach for Parallel Corpora Based Cross Language Information Retrieval. In *ECAI'12*. 79–84.

[2] Hosein Azarbonyad, Azadeh Shakery, and Heshaam Faili. 2019. A learning to rank approach for cross-language information retrieval exploiting multiple translation resources. *Natural Language Engineering* 25, 3 (2019), 363–384.

[3] Adam Berger and John Lafferty. 1999. Information Retrieval As Statistical Translation. In *SIGIR '99*. 222–229.

[4] Chris Buckley, Mandar Mitra, Janet Walz, and Claire Cardie. 2000. Using Clustering and SuperConcepts Within SMART: TREC 6. *Inf. Process. Manage.* 36, 1 (Jan. 2000), 109–131.

[5] Kareem Darwish and Douglas W. Oard. 2003. Probabilistic Structured Query Methods. In *SIGIR '03*. 338–344.

[6] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A Formal Study of Information Retrieval Heuristics. In *SIGIR '04*. 49–56.

[7] Elham Ghanbari and Azadeh Shakery. 2019. Query-dependent Learning to Rank for Cross-lingual Information Retrieval. *Knowl. Inf. Syst.* 59, 3 (June 2019), 711–743.

[8] Daqing He, Douglas W. Oard, Jianqiang Wang, Jun Luo, Dina Demner-Fushman, Kareem Darwish, Philip Resnik, Sanjeev Khudanpur, Michael Nossal, Michael Subotin, and Anton Leuski. 2003. Making MIRACLEs: Interactive Translingual Search for Cebuano and Hindi. *ACM Trans. Asian Lang. Inf. Process.* 2, 3 (Sept. 2003), 219–244.

[9] Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit*. 79–86.

[10] Wessel Kraaij and Franciska de Jong. 2004. Transitive Probabilistic CLIR Models. In *RIAO '04*. 69–81.

[11] Wessel Kraaij, Jian-Yun Nie, and Michel Simard. 2003. Embedding Web-based Statistical Translation Models in Cross-language Information Retrieval. *Comput. Linguist.* 29, 3 (Sept. 2003), 381–419.

[12] Victor Lavrenko, Martin Choquette, and W. Bruce Croft. 2002. Cross-lingual Relevance Models *(SIGIR '02)*. 175–182.

[13] Bo Li and Eric Gaussier. 2012. An Information-based Cross-language Information Retrieval Model *(ECIR'12)*. Springer-Verlag, Berlin, Heidelberg, 281–292.

[14] Paul Mcnamee and James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *Inf. Retr.* 7, 1-2 (Jan. 2004), 73–97.

[15] Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*. Morgan and Claypool Publishers.

[16] Douglas W. Oard and Jianqiang Wang. 2001. NTCIR-2 ECIR Experiments at Maryland: Comparing Pirkola's Structured Queries and Balanced Translation. In *NTCIR-2*.

[17] Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Comput. Linguist.* 29, 1 (March 2003), 19–51.

[18] Razieh Rahimi and Azadeh Shakery. 2017. Online Learning to Rank for Cross-Language Information Retrieval *(SIGIR '17)*. 1033–1036.

[19] Razieh Rahimi, Azadeh Shakery, and Irwin King. 2014. Axiomatic Analysis of Cross-Language Information Retrieval *(CIKM '14)*. 1875–1878.

[20] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIRâĂŹ94*. Springer, 232–241.

[21] Jacques Savoy. 2005. Comparative Study of Monolingual and Multilingual Search Models for Use with Asian Languages. *ACM Trans. Asian Lang. Inf. Process.* 4, 2 (June 2005), 163–189.

[22] P. Sorg and P. Cimiano. 2012. Exploiting Wikipedia for Cross-lingual and Multilingual Information Retrieval. *Data Knowl. Eng.* 74 (April 2012), 26–45.

[23] Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. In *SIGIR '15*. 363–372.

[24] Jinxi Xu, Ralph Weischedel, and Chanh Nguyen. 2001. Evaluating a Probabilistic Model for Cross-lingual Information Retrieval. In *SIGIR '01*. 105–110.