An Information-Based Cross-Language Information Retrieval Model

Bo Li, Éric Gaussier

An Information-based Cross-Language Information Retrieval Model

Bo Li and Eric Gaussier

Université J. Fourier-Grenoble 1/CNRS - Laboratoire d'Informatique de Grenoble (LIG)
firstname.lastname@imag.fr

Abstract. We present in this paper well-founded cross-language extensions of the recently introduced models in the information-based family for information retrieval, namely the LL (log-logistic) and SPL (smoothed power law) models of [4]. These extensions are based on (a) a generalization of the notion of information used in the information-based family, (b) a generalization of the random variables also used in this family, and (c) the direct expansion of query terms with their translations. We then review these extensions from a theoretical point-of-view, prior to assessing them experimentally. The results of the experimental comparisons between these extensions and existing CLIR systems, on three collections and three language pairs, reveal that the cross-language extension of the LL model provides a state-of-the-art CLIR system, yielding the best performance overall.

1 Introduction

Cross-Language Information Retrieval (CLIR) is concerned with the problem of finding documents written in a language different from that of the query. If attempts to model multilinguality in information retrieval date back from the early seventies [15], a renewed interest was brought to the field by the rise of the Web in the mid-nineties, as pages written in many different languages were all of a sudden availability. International organizations, governments of multi-lingual countries, to name the most important, have been traditional users of CLIR systems, and the need for such systems in everyday life becomes more and more clear, with the development of travels and tourism, to name but a few (the recent book by J.-Y. Nie on Cross-Language Information Retrieval [11] exposes in detail the need for cross-language and multilingual IR).

There are several ways to cross the language barrier in CLIR models: through mapping the document representation into the query representation space (an approach known as document translation), through mapping the query representation into the document representation (an approach known as query translation) or through mapping both representations into a third space (interlingua approach). As for implementation, existing CLIR models fall into two categories: model-independent approaches and model-dependent approaches. Model-independent approaches treat translation and retrieval as two separate processes. The queries or the documents are first translated into the corresponding language of the documents or the queries. Monolingual IR models are then applied directly. A typical and also broadly used approach of this type is the machine translation (MT) approach (e.g. [10, 3]) which employs MT systems to

translate the queries or documents before the monolingual retrieval process. Model-dependent methods integrate the translation and retrieval processes in a uniform framework. These methods, developed in e.g. [9, 10] in the context of language models, have the advantage of accounting better for the uncertainty of translation during retrieval.

Most model-dependent approaches to CLIR rely on a cross-language extension of existing monolingual IR systems, as the ones we have already mentioned for the language modeling approach. If most monolingual IR systems have been extended to a cross-language setting, it is not true for all of them, and we explore in this paper the cross-language extension of the recently introduced information-based family of IR systems. Two models in this family have been shown to provide state-of-the-art performance in monolingual IR, and the question of their possible extension, and the quality of this extension, to a cross-language setting remains unanswered.

The remainder of the paper is organized as follows: Section 2 first introduces the family of information-based models for IR prior to presenting three possible extensions of two models in this family to the cross-language setting; it also introduces a theoretical validation of cross-language IR models through a condition such models should satisfy and termed the Dilution/Concentration condition. Section 3 then presents the experimental validation, assessing the different extensions presented in Section 2, and comparing the CLIR systems introduced with existing ones on three collections and three language pairs. Lastly, Section 4 concludes the paper. The notations we use throughout the paper are summarized in Table 1 (w represents a word).

Table 1. Notations used in the paper

Notation	Description
x_w^q	Number of occurrences of w in query q
x_w^d	Number of occurrences of w in document d
t_w^d	Normalized version of x_w^d
l_d	Length of document d
l_m	Average document length
L	Length of document collection
N	Number of documents in the collection
N_w	Number of documents containing w
$TS(w)$	Set of translations of w
$DS(w)$	Set of documents containing w ($N_w = \|DS(w)\|$)
RSV(q, d)	Retrieval Status Value of doc. d for query q

2 Information-Based Models

Information-based models for IR, recently introduced in [4], compute the similarity between queries and documents through the quantity of information brought by document terms on query words. Two such models, referred to as Log-Logistic model (in short LL) and Smoothed Power Law model (in short SPL), were shown in [4, 5] to be either on par or to outperform other IR models on several collections and in different settings, as

the one of pseudo-relevance feedback. We want here to explore possible cross-language extensions of these models and to assess their behavior in a CLIR setting.

Information-based models are based on the following retrieval status value[1]:

$$RSV(q,d) = \sum_{w \in q} -\frac{x_w^q}{l_q} \log P(X_w \geq t_w^d | \lambda_w)$$
$$= \sum_{w \in q \cap d} -\frac{x_w^q}{l_q} \log P(X_w \geq t_w^d | \lambda_w) \quad (1)$$

where:

- t_w^d is a normalization function depending on the number of occurrences, x_w^d, of w in d, and on the length, l_d, of d, and satisfies:

$$\frac{\partial t_w^d}{\partial x_w^d} > 0; \quad \frac{\partial t_w^d}{\partial l_d} < 0; \frac{\partial^2 x_w^d}{\partial (t_w^d)^2} \geq 0$$

  In this work, and following [4], it is defined as: $t_w^d = x_w^d \log(1 + c\frac{l_m}{l_d})$ where c is the smoothing parameter;
- P is a probability distribution defined over a random variable, X_w, associated to each word w. This probability distribution must be:
  - Continuous, the random variable under consideration, t_w^d, being continuous;
  - Compatible with the domain of t_w^d, i.e. if t_{min} is the minimum value of t_w^d, then $P(X_w \geq t_{\text{min}} | \lambda_w) = 1$;
  - Bursty, i.e. it should be such that:
    $\forall \epsilon > 0,\ g_\epsilon(x) = P(X \geq x + \epsilon | X \geq x)$ is strictly increasing in x;
- And λ_w is a collection-dependent parameter of P. As suggested in [4], it is set as:

$$\lambda_w = \frac{N_w}{N} \quad (2)$$

As one can note, equation 1 computes the information brought by the document on each query word ($-\log P(X_w \geq t_w^d | \lambda_w)$) weighted by the importance of the word in the query ($\frac{x_w^q}{l_q}$). In order to define a proper IR model, one needs to choose a particular bursty distribution. As mentioned above, two such distributions have been proposed and studied, and we will rely on them here. These are the log-logistic and smoothed power law distributions, associated to the models referred to as LL and SPL and defined as (see [4]):

$$RSV_{LL}(q,d) = \sum_{w \in q \cap d} -\frac{x_w^q}{l_q} \log \frac{\lambda_w}{\lambda_w + t_w^d}$$

$$RSV_{SPL}(q,d) = \sum_{w \in q \cap d} -\frac{x_w^q}{l_q} \log \frac{\lambda_w^{\frac{t_w^d}{t_w^d+1}} - \lambda_w}{1 - \lambda_w}$$

We now turn to cross-language extensions of this family of models.

[1] We introduce a slight modification, namely the normalization by the query length, in the formula given in [4], in order to provide a more intuitive explanation of the models. This modification does not change the ranking of the documents.

2.1 Cross-Language Extensions

First of all, one can note that the information brought by a document on a query term in equation 1 is restricted to the query word itself. It is however possible to adopt a more general view by considering the mean information brought by all words in the document related to a given query term. Let $\mathcal{F}(w)$ denote the set of all the words related, through a relation we leave unspecified for the moment, to a given word w. Let us furthermore introduce the normalized relation between w and a word w' in d, a quantity we will denote as $\mathcal{A}(w, w', d)$, as:

$$\mathcal{A}(w, w', d) = \begin{cases} \frac{I_{\mathcal{F}(w)}(w')}{\sum_{w'' \in d} I_{\mathcal{F}(w)}(w'')} & \text{if } \sum_{w'' \in d} I_{\mathcal{F}(w)}(w'') > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $I_{\mathcal{F}(w)}$ is the indicator function of the set $\mathcal{F}(w)$. The mean information brought by all words of the document d related to a given query term w can then be defined as: $-\sum_{w' \in d} \mathcal{A}(w, w', d) \log P(X_{w'} \geq t^d_{w'} | \lambda_{w'})$, leading to the overall retrieval function:

$$RSV(q, d) = -\sum_{w \in q} \frac{x^q_w}{l_q} \sum_{w' \in d} \mathcal{A}(w, w', d) \log P(X_{w'} \geq t^d_{w'} | \lambda_{w'})$$

Equation 1 is just a special case of the above formulation, obtained by setting $\mathcal{F}(w) = \{w\}$, i.e. considering that words are only related to themselves. The application to a cross-language setting then simply amounts to using the translation relation, $\mathcal{F}(w) = TS(w)$, for computing $\mathcal{A}(w, w', d)$ ($TS(w)$ denotes the translation set of word w in our general notations). This leads, for the LL and SPL models, to:

$$RSV_{LL}(q, d) = -\sum_{w \in q} \frac{x^q_w}{l_q} \sum_{w' \in d} \mathcal{A}(w, w', d) \log(\frac{\lambda_{w'}}{\lambda_{w'} + t^d_{w'}}) \quad (3)$$

$$RSV_{SPL}(q, d) = -\sum_{w \in q} \frac{x^q_w}{l_q} \sum_{w' \in d} \mathcal{A}(w, w', d) \log(\frac{\lambda_{w'}^{\frac{t^d_{w'}}{t^d_{w'}+1}} - \lambda_{w'}}{1 - \lambda_{w'}}) \quad (4)$$

The above equations define two new CLIR models, which we will refer to as MI_{LL} and MI_{SPL}, MI standing for *Mean Information*.

A second extension consists in considering that the random variable used in the information-based family is not associated to a single word w, but to a set of words $\mathcal{F}(w)$, namely the words related to w. This defines a new retrieval function of the form:

$$RSV(q, d) = -\sum_{w \in q} \frac{x^q_w}{l_q} \log P(X_{\mathcal{F}(w)} \geq t^d_{\mathcal{F}(w)} | \lambda_{\mathcal{F}(w)})$$

As before, equation 1 is just a special case obtained by setting $\mathcal{F}(w) = \{w\}$, and a cross-language version can be obtained by setting $\mathcal{F}(w) = TS(w)$. One needs however to define $t^d_{\mathcal{F}(w)}$ and $\lambda_{\mathcal{F}(w)}$. We simply set here the first quantity to the sum of the corresponding quantities for the words in $\mathcal{F}(w)$, which corresponds to the fact that we

have indeed observed that many (normalized) occurrences of w in d, through its related words. The second quantity is set in a similar fashion, by considering the normalized document frequency of all the words in $\mathcal{F}(w)$ (see equation 2). This leads to the following cross-language version of the LL and SPL models:

$$t^d_{\mathcal{F}(w)} = \sum_{w' \in \mathcal{F}(w)} t^d_{w'}$$

$$\lambda_{\mathcal{F}(w)} = \frac{|\cup_{w' \in \mathcal{F}(w)} DS(w')|}{N}$$

$$\text{RSV}_{LL}(q,d) = -\sum_{w \in q} \frac{x^q_w}{l_q} \log(\frac{\lambda_{\mathcal{F}(w)}}{t^d_{\mathcal{F}(w)} + \lambda_{\mathcal{F}(w)}}) \tag{5}$$

$$\text{RSV}_{SPL}(q,d) = -\sum_{w_f \in q} \frac{x^q_w}{l_q} \log(\frac{(\lambda_{\mathcal{F}(w)})^{\frac{t^d_{\mathcal{F}(w)}}{t^d_{\mathcal{F}(w)}+1}} - \lambda_{\mathcal{F}(w)}}{1 - \lambda_{\mathcal{F}(w)}}) \tag{6}$$

The above extension bears strong similarities with the SYN operator of the INQUERY system, developed for CLIR purposes in [12]. Indeed, the above formulation can also be obtained by considering that all related words form a single word. We have shown here, however, that it also derives from a different perspective, through the use, in the information-based family, of a single random variable to account for all related words. The setting of the associated parameters (t^d_w and λ_w) then naturally follows from the general framework of the information-based family of IR models. For this reason, we will refer to the above CLIR models as JV$_{LL}$ and JV$_{SPL}$, JV standing for *Joint random Variable*.

Lastly, a third cross-language extension can directly be obtained by expanding all query terms with their translations. As in standard bilingual dictionaries translations are not weighted, we resort to the following, simple extension of equation 1, which could however be extended by taking translation weights into account:

$$RSV(q,d) = -\sum_{w \in q} \frac{x^q_w}{l_q} \sum_{w' \in d \cap TS(w)} \log P(X_{w'} \geq t^d_{w'} | \lambda_{w'})$$

This leads, for the LL and SPL models, to:

$$RSV_{LL}(q,d) = -\sum_{w \in q} \frac{x^q_w}{l_q} \sum_{w' \in d \cap TS(w)} \log(\frac{\lambda_{w'}}{\lambda_{w'} + t^d_{w'}}) \tag{7}$$

$$RSV_{SPL}(q,d) = -\sum_{w \in q} \frac{x^q_w}{l_q} \sum_{w' \in d \cap TS(w)} \log(\frac{\lambda_{w'}^{\frac{t^d_{w'}}{t^d_{w'}+1}} - \lambda_{w'}}{1 - \lambda_{w'}}) \tag{8}$$

We will refer to the above CLIR models as QE$_{LL}$ and QE$_{SPL}$, QE standing for *Query Expansion*.

To summarize, we have defined, through the above developments, three new CLIR versions of the LL and SPL models, within the general framework of information-based models for IR:

1. MI_{LL} and MI_{SPL}, corresponding to equations 3 and 4;
2. JV_{LL} and JV_{SPL}, corresponding to equations 5 and 6;
3. QE_{LL} and QE_{SPL}, corresponding to equations 7 and 8.

Prior to turning to the experimental validation of these models, we want to address the question of whether it is possible to validate them from a theoretical point of view. We do so in the following section by resorting to the axiomatic theory of IR.

2.2 Theoretical Validation

Heuristic retrieval constraints were first fully developed in the seminal work of Fang et al. [7], and followed by many others since, including [8, 6, 4, 5, 17]. Such constraints state conditions IR models should satisfy, and there is now a large corpus of empirical evidence showing that models failing on one condition do not yield an optimal performance. As shown in [4], the LL and SPL models we have considered comply with all the conditions for *ad hoc* information retrieval, and so do their cross-language extensions. However, the cross-language setting also relies on new elements, the translations, and the question remains as to whether these new elements can be regulated through a particular CLIR condition. We develop such a condition below.

Let us assume a collection of French documents about rivers and lakes, and the English query *bank*. In this context, the possible translations of *bank* in French are *rive, berge, banc*[2]. Now let us assume that, in one document d, the words *berge* and *banc* appears two times each, and that, in another document, d', the word *rive* appears four times. Let us also assume that d and d' have roughly the same length and that *berge* and *banc* only occur in d and *rive* only in d'. All these assumptions can be met, for example, on a collection containing formatted articles on water flows. In this context, there is absolutely no difference between d and d' with respect to their relevance according to the query, and one would like a good CLIR strategy to assign the same score to these two documents. The following condition formalizes this.

Condition 1 *Let q be a source language query consisting of a single term w, d and d' two target language documents of equal length. Furthermore, let $\{w'_0, w'_1, \cdots, w'_k\}$ be equally likely and equally good translations of w such that:*

$$\begin{cases} x^d_{w'_i} = 1, N_{w'_i} = 1, \ 1 \leq i \leq k \\ x^{d'}_{w'_0} = k, N_{w'_0} = 1 \end{cases}$$

Then, a good CLIR strategy should satisfy:

$$RSV(q, d) = RSV(q, d')$$

[2] One can certainly think of other possible translations, but this does not change our argument.

Because the translation of w is either diluted, in d, on several words, or concentrated, in d', on a single word, we will refer to the above condition as the *DC condition*, where DC stands for Dilution/Concentration. We now review the different CLIR models we proposed in light of this condition, focusing here on the LL model, the reasoning and results being the same for SPL.

As all translations in d have the same number of occurrences, they also have the same normalized frequency, which will be denoted by τ: $t^d_{w'_i} = \tau,\ 1 \leq i \leq k$. We furthermore have: $t^{d'}_{w'_0} = k\tau$. The DC assumptions furthermore imply that all translations have the same parameter λ: $\lambda_{w'_i} = \frac{1}{N},\ 0 \leq i \leq k$. Given this, we have:

- For the QE extension:

  $$RSV_{LL}(q,d) = k\log(\tau N + 1),\ RSV_{LL}(q,d') = \log(k\tau N + 1)$$

  The function $RSV_{LL}(q,d') - RSV_{LL}(q,d)$ is strictly decreasing with τ, its derivative being strictly negative, and equals 0 when $\tau = 0$, which implies that:

  $$RSV_{LL}(q,d') < RSV_{LL}(q,d)$$

  The QE strategy thus does not fulfill the DC condition.
- For the MI extension:

  $$RSV_{LL}(q,d) = \log(\tau N + 1),\ RSV_{LL}(q,d') = \log(k\tau N + 1)$$

  This time, the function $RSV_{LL}(q,d') - RSV_{LL}(q,d)$ is increasing with τ for $k \geq 1$, and equals 0 when $\tau = 0$, which implies that:

  $$RSV_{LL}(q,d') > RSV_{LL}(q,d)$$

  The MI strategy thus does not fulfill the DC condition. One can note however that in this extension $\text{RSV}(q,d)$ is closer to $\text{RSV}(q,d')$ than in the QE one. Indeed, let us denote by RSV_{QE} and RSV_{MI} the different retrieval functions associated with the different extensions. We have: $\text{RSV}_{QE}(q,d') = \text{RSV}_{MI}(q,d') = \text{RSV}(q,d')$. The function $\text{RSV}_{QE}(q,d) + \text{RSV}_{MI}(q,d) - 2\text{RSV}(q,d')$ is increasing with τ (its derivative being positive as soon as $k\tau N > 1$, which is the case in practice) and equals 0 when $\tau = 0$. Hence:

  $$\text{RSV}_{QE}(q,d) - \text{RSV}_{QE}(q,d') > \text{RSV}_{MI}(q,d') - \text{RSV}_{MI}(q,d)$$

- For the JV extension: $RSV_{JV}(q,d) = \log(k\tau N + 1) = RSV_{JV}(q,d')$. This extension is thus fully compliant with the DC condition.

The above theoretical development thus reveals that both the MI and QE extensions do not fulfill the DC condition, the violation of the condition being less marked in the MI extension. Furthermore, the JV extension does fulfill the DC condition. As we will see, our experiments are in agreement with these findings.

3 Experimental Validation

We use in our experiments the English collections from the bilingual tasks of the CLEF campaign[3], with English, French, German and Italian queries, from the year 2000 to 2004. Table 2 lists the number of documents (N_d), number of distinct words (N_w), average document length (DL_{avg}) in the English document collections, as well as the number of queries, N_q, in each task (all the queries are available in all languages). As the queries from the year 2000 to 2002 have the same target collection, they are combined in a single task. In all our experiments, we use bilingual dictionaries comprising respectively 70k entries for the French-English language pair, 58k entries for the German-English language pair, and 67k entries for the Italian-English language pair. For evaluation, we use the Mean Average Precision (MAP) scores to evaluate the different models. Lastly, we rely on a paired t-test (at the level 0.05) to measure significance difference between the different systems.

Table 2. Characteristics of different CLEF collections

Collection	N_d	N_w	DL_{avg}	N_q
CLEF 2000-2002	113,005	173,228	310.85	140
CLEF 2003	169,477	232,685	284.09	60
CLEF 2004	56,472	119,548	230.52	50

3.1 Validation of the Information-based Extensions

In a first series of experiments, we compare the different extensions (MI, JV and QE) proposed for both the LL and SPL models. Information-based models rely on one parameter, namely c, used in the normalization step. As this normalization step is identical to the one used in DFR models ([1]), we use the default setting provided in Terrier[4] for this parameter: $c = 1$. The results we obtained for MAP scores are displayed, for the three language pairs (i.e. French(Fr)-English(En), Italian(It)-English(En), and German(De)-English(En)), in Table 3. As one can note, and in accordance to the theoretical validation developed in section 2.2, for both LL and SPL, the JV extension is significantly better than both the MI and QE extensions, and meanwhile MI provides better results than QE. In the following experiments, aimed at comparing different CLIR systems, we will thus only rely on the JV extension for the two models LL and SPL of the information-based family.

3.2 Comparison with other CLIR Models

We compare now the cross-language versions of LL and SPL we have introduced with CLIR versions of standard systems, namely: (a) a vector space model based on Robertson's *tf* and Sparck Jones' *idf* ([14]), referred to as TF-IDF, (b) BM25 with the default

[3] http://www.clef-campaign.org

[4] terrier.org

Table 3. Comparison of different cross-language extensions of LL and SPL in terms of MAP scores. A † indicates, for each model, that the difference with the best performing extension (in bold) is significant.

Collection		LL			SPL		
		JV	QE	MI	JV	QE	MI
CLEF 2000-2002	Fr-En	**0.4174**	0.2042†	0.3748†	0.4008†	0.1937†	0.3702†
	It-En	**0.3934**	0.2117†	0.3704†	0.3730†	0.1844†	0.3417†
	De-En	**0.4102**	0.2124†	0.3750†	0.3901†	0.1990†	0.3574†
CLEF 2003	Fr-En	**0.4801**	0.2229†	0.4167†	0.4615†	0.2039†	0.4201†
	It-En	**0.4339**	0.2133†	0.3817†	0.4210†	0.1991†	0.3746†
	De-En	**0.4625**	0.2200v	0.3942†	0.4438†	0.2032†	0.3277†
CLEF 2004	Fr-En	**0.5204**	0.3085†	0.4171†	0.4317†	0.2317†	0.3460†
	It-En	**0.4910**	0.2973†	0.4058†	0.4213†	0.2087†	0.3170†
	De-En	**0.4921**	0.2969†	0.4062†	0.4222†	0.2166†	0.3276†

parameter setting given by the Terrier system, (c) INQUERY with the default parameters of the Lemur system[5] and (d) the Jelinek-Mercer and Dirichlet versions of the language models, again with the default parameters of the Terrier system ($\lambda = 0.15$ and $\mu = 2500$), referred to as LM-JM and LM-DIR. For the first three models, we directly rely on the SYN strategy, which amounts to considering all the translations of a given query term in the documents as forming a single word. This strategy has been shown to outperform other ones in different studies ([12, 16, 13, 2]). For LM-JM and LM-DIR, two additional strategies have been explored in previous studies (e.g. [10]): integration of the translations within the query model (QT), or within the document model (DT), and we first compare them here.

The results of the comparison between the three LM-related strategies (SYN, QT, DT) are given in Table 4, for the MAP scores on three language pairs. As one can note, the SYN strategy outperforms the other ones, the difference being always significant. Because of that, we will rely for LM-JM and LM-DIR on the SYN strategy in the following experiments.

It is also interesting to note that DT yields results consistently better than QT, which is the worst performing strategy. Interestingly, QT is the only strategy which does not fulfill the DC condition introduced in section 2.2. Indeed, for Jelinek-Mercer smoothing with the smoothing parameter λ (the reasoning and the results are the same for Dirichlet smoothing), we obtain, under the setting of the DC condition[6]:

$$\text{RSV}_{QT}(q, d') - \text{RSV}_{QT}(q, d) = \alpha(\log k - (k-1)\log((1-\lambda)\frac{1}{l_d} + \lambda\frac{1}{L}))$$

where α corresponds to the translation probability between the query word and any of its translation. The above quantity is strictly positive for $k \geq 1$. In contrast, both the DT and SYN strategy are compliant with the DC condition. It is straightforward to see this for SYN: the different words in d are grouped into a single word with k occurrences,

[5] www.lemurproject.org

[6] We omit the derivation here as it is direct and purely technical.

hence making the setting in d identical to the one in d'. For DT, we obtain:

$$\text{RSV}_{DT}(q, d') = \log\left(k\alpha((1-\lambda)\frac{1}{l_d} + \lambda\frac{1}{L})\right) = \text{RSV}_{DT}(q, d)$$

Table 4. Comparison of different CLIR strategies (SYN, QT, DT) for language models in terms of MAP scores. A † indicates, for each model, that the difference with the best performing extension (in bold) is significant. For clarity sake, when the difference with the best result is not significant, the result is italicized.

Collection		DT		QT		SYN	
		JM	DIR	JM	DIR	JM	DIR
CLEF 2000-2002	Fr-En	0.3711†	0.3924†	0.3641†	0.3491†	0.3930†	**0.4102**
	It-En	0.3497†	0.3660†	0.3207†	0.3143†	0.3720†	**0.3878**
	De-En	0.3728†	0.3797†	0.3490†	0.3504†	*0.3925*	**0.3983**
CLEF 2003	Fr-En	0.4419†	0.4038†	0.3981†	0.3781†	**0.4716**	0.4242†
	It-En	0.4162†	0.4211†	0.3745†	0.3801†	**0.4355**	0.3857†
	De-En	0.4271†	0.3713†	0.3813†	0.3336†	**0.4554**	0.4098†
CLEF 2004	Fr-En	0.4217†	0.4222†	0.3861†	0.4186†	**0.4513**	*0.4417*
	It-En	0.3907†	0.3824†	0.3778†	0.3812†	**0.4221**	*0.4201*
	De-En	0.3992†	0.3874†	0.3810†	0.3796†	**0.4331**	*0.4310*

Lastly, Table 5 gives the results obtained with the different CLIR systems we have reviewed, on all language pairs and all collections. The performance of the monolingual version of the CLIR systems is given in the line MON. First of all, one can note that either the model LL_{JV} obtains the best score (9 times out of 12) , or the difference with the best system is not significant. Furthermore, when LL_{JV} obtains the best score, the difference with the other models is most of the time significant. Indeed, for the cross-language part, only LM-DIR is on a par with LL_{JV} on the 2000-2002 collection, only LM-JL is on a par with LL_{JV} on the 2003 collection, and all models are significantly below LL_{JV} on the 2004 collection.

4 Conclusion

Several previous studies were based on the cross-language strategies we have reviewed here for embedding dictionaries (either manually or automatically built) in CLIR systems. None of them however addressed the problem of extending the recently introduced information-based models to a cross-language setting. We have presented here several possible strategies for such an extension, through the generalization of the information used in information-based models, through the generalization of the random variables also used in this family, and through the expansion of query terms. The strategy based on the generalization of the random variables play a role similar to the one of the SYN strategy reviewed in previous studies. The good behavior of this strategy, noticed in these previous studies, is confirmed here for the information-based family.

Table 5. Comparison of different CLIR systems in terms of MAP scores on all language pairs and collections. A † indicates, for each model, that the difference with the best performing extension (in bold) is significant. For clarity sake, when the difference with the best result is not significant, the result is italicized.

Data	Model	TF-IDF	BM25	LM-JM	LM-DIR	INQUERY	LL_{JV}	SPL_{JV}
CLEF 2000-2002	MON	0.4475†	0.4744†	0.4621†	0.4783†	0.4227†	**0.4866**	*0.4828*
	Fr-En	0.3641†	0.3891†	0.3990†	*0.4102*	0.3527†	**0.4174**	0.4008†
	It-En	0.3325†	0.3578†	0.3720†	*0.3878*	0.3216†	**0.3934**	0.3730†
	Ge-En	0.3502†	0.3674†	*0.3925*	*0.3983*	0.3419†	**0.4102**	0.3901†
CLEF 2003	MON	0.4763†	**0.5031**	0.4919†	0.4751†	0.4369†	*0.5030*	*0.5001*
	Fr-En	0.4155†	0.4405†	*0.4716*	0.4242†	0.4076†	**0.4801**	0.4615†
	It-En	0.3764†	0.4000†	**0.4355**	0.3857†	0.3732†	*0.4339*	0.4210†
	Ge-En	0.3966†	0.4198†	*0.4554*	0.4098†	0.3842†	**0.4625**	0.4438†
CLEF 2004	MON	0.5187†	0.5228†	0.5110†	**0.5386**	0.4264†	*0.5381*	0.5290†
	Fr-En	0.4225†	0.4197†	0.4513†	0.4417†	0.3763†	**0.5204**	0.4317†
	It-En	0.3917†	0.3834†	0.4221†	0.4201†	0.3425†	**0.4910**	0.4213†
	Ge-En	0.4008†	0.3947†	0.4331†	0.4310†	0.3534†	**0.4921**	0.4222†

We have furthermore introduced a new CLIR condition, thus extending the axiomatic approach to IR to the cross-language setting. This new condition, which we referred to as the Dilution/Concentration condition, helped us assess from a purely theoretical point-of-view the different cross-language extensions we have introduced. The results obtained from this theoretical assessment were confirmed in our experiments. We also used this condition to assess the possible strategies for building cross-language extensions in the language modeling approach to IR, and again found that the theoretical results are in line with the experimental ones.

Lastly, we have shown that the cross-language extension of the log-logistic model (LL) based on the joint random variable (and equivalent to the SYN strategy) yields the best performance on three collections and three language pairs. This model is never significantly below any other model, always significantly above most of them if not all of them. We thus believe this model to be a state-of-the-art CLIR model. Its simple form, given by equation 5, also makes it appealing from an implementation perspective.

In the future, we plan on exploring different settings of the parameters of each model. We have relied in our experiments on the default setting recommended in different IR systems (Lemur and Terrier), as this is a setting commonly used in cross-language information retrieval when new collections and queries have to be processed. This is also the setting used by many participants to cross-language evaluation campaigns who just want to rely on black-box CLIR systems and develop pre-processing or post-processing components.

References

1. Amati, G., Van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst. 20, 357–389 (October 2002)

2. Ballesteros, L., Sanderson, M.: Addressing the lack of direct translation resources for cross-language retrieval. In: Proceedings of the twelfth international conference on Information and knowledge management. pp. 147–152. CIKM '03, New Orleans, LA, USA (2003)
3. Braschler, M.: Combination approaches for multilingual text retrieval. Inf. Retr. 7, 183–204 (January 2004)
4. Clinchant, S., Gaussier, E.: Information-based models for ad hoc ir. In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 234–241. SIGIR '10, ACM, New York, NY, USA (2010)
5. Clinchant, S., Gaussier, É.: Is document frequency important for prf? In: ICTIR. pp. 89–100 (2011)
6. Cummins, R., O'Riordan, C.: An axiomatic comparison of learned term-weighting schemes in information retrieval: clarifications and extensions. Artif. Intell. Rev. 28, 51–68 (June 2007)
7. Fang, H., Tao, T., Zhai, C.: A formal study of information retrieval heuristics. In: SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (2004)
8. Fang, H., Zhai, C.: 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. pp. 115–122. SIGIR '06 (2006)
9. Federico, M., Bertoldi, N.: Statistical cross-language information retrieval using n-best query translations. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 167–174. SIGIR '02, Tampere, Finland (2002)
10. Kraaij, W., Nie, J.Y., Simard, M.: Embedding web-based statistical translation models in cross-language information retrieval. Computational Linguistic 29, 381–419 (September 2003)
11. Nie, J.Y.: Cross-Language Information Retrieval. Morgan & Claypool, New York, NY, USA (2010)
12. Pirkola, A.: The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 55–63. SIGIR '98, ACM, New York, NY, USA (1998)
13. Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K.: Dictionary-based cross-language information retrieval: Problems, methods, and research findings. Inf. Retr. 4, 209–230 (September 2001)
14. Robertson, S.E., Sparck Jones, K.: Relevance weighting of search terms, pp. 143–160 (1988)
15. Salton, G.: Automatic processing of foreign language documents. In: Proceedings of the 1969 conference on Computational linguistics. pp. 1–28. COLING '69, Association for Computational Linguistics, Stroudsburg, PA, USA (1969)
16. Sperer, R., Oard, D.W.: Structured translation for cross-language information retrieval. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 120–127. SIGIR '00, Athens, Greece (2000)
17. Zhai, C.: Axiomatic analysis and optimization of information retrieval models. In: ICTIR (2011)