C. Peters et al. (Eds.): Proceedings of CLEF 2009, Corfu, Greece
Multilingual Information Access Evaluation Vol. I: Text Retrieval Experiments
LNCS, ©Springer 2010.

Evaluating Cross-Language Explicit Semantic Analysis and Cross Querying

Maik Anderka, Nedim Lipka, and Benno Stein

Bauhaus-Universität Weimar, 99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

Abstract This paper describes our participation in the TEL@CLEF task of the CLEF 2009 ad-hoc track. The task is to retrieve items from various multilingual collections of library catalog records, which are relevant to a user's query. Two different strategies are employed: (*i*) the Cross-Language Explicit Semantic Analysis, CL-ESA, where the library catalog records and the queries are represented in a multilingual concept space that is spanned by aligned Wikipedia articles, and, (*ii*) a Cross Querying approach, where a query is translated into all target languages using Google Translate and where the obtained rankings are combined. The evaluation shows that both strategies outperform the monolingual baseline and achieve comparable results.
Furthermore, inspired by the Generalized Vector Space Model we present a formal definition and an alternative interpretation of the CL-ESA model. This interpretation is interesting for real-world retrieval applications since it reveals how the computational effort for CL-ESA can be shifted from the query phase to a preprocessing phase.

1 Introduction

Cross-language information retrieval, CLIR, is the task of retrieving documents from a target collection written in a language different from the language of a user's query. CLIR systems give multilingual users the possibility to express queries in any language, e.g., their native language, and to obtain result documents in all languages they are familiar with. Since CLIR is not restricted to collections in the query language more sources can be included in the retrieval process, and the chance to fulfill a particular information need of a multilingual user is higher. Another use case for CLIR techniques is cross-language plagiarism detection, where the query corresponds to a suspicious document and the target collection is a reference corpus with original documents [3].

The Cross-Language Evaluation Forum, CLEF, provides an infrastructure for the evaluation of information retrieval systems, both monolingual and cross-lingual. We participated in the TEL@CLEF task of the CLEF 2009 ad-hoc track, which aims at the evaluation of systems to retrieve relevant items from multilingual collections of library catalog records. The main challenges of this task are the multilinguality and the sparsity of the dataset. We used two different CLIR approaches to tackle this task; the paper in hand outlines and discusses these approaches and the achieved results.

The first approach is Cross-Language Explicit Semantic Analysis, CL-ESA, which is a multilingual retrieval model to access cross-language similarity between text documents [3]. The CL-ESA model exploits a document-aligned comparable corpus such as Wikipedia in order to map the query and the documents into a common multilingual concept space [3,4]. We also present a formal definition and an alternative interpretation for the CL-ESA model, which is inspired by the Generalized Vector Space Model, GVSM. Our view is mathematically equivalent to the original idea of the CL-ESA model; it reveals how the computational effort for CL-ESA can be shifted from the query phase to a preprocessing phase.

In the second approach, called Cross Querying, each query is translated into all target languages. The particular rankings are used in a combined fashion considering the most likely language of the documents. The evaluation on the TEL@CLEF collections shows that both CLIR approaches are able to outperform the monolingual baseline. In the bilingual subtask, querying with a foreign language, Cross Querying achieves nearly the same or even higher results compared to the monolingual subtask; the performance of the CL-ESA is lower compared to the monolingual results.

The paper is organized as follows. Section 2 describes the target collection used in the TEL@CLEF task along with the evaluation procedure. Section 3 defines the general CL-ESA model, our formalization, and details of the CL-ESA implementation employed in the experiments. Section 4 presents the Cross Querying approach, Section 5 discusses the evaluation, and Section 6 concludes with an outlook.

2 TEL@CLEF Dataset and Evaluation Procedure

In this year's TEL@CLEF task three target collections, provided by The European Library, TEL, are used. The collections are labeled BL, ONB, and BNF, and mainly contain information in English, German, and French respectively (see Table 1). The collections are comprised of library catalog records, referring to different types of items such as articles, books, or videos. The data is provided in structured form and represented in XML. Each library catalog record has several fields containing meta information and content information that describe the particular item. Typical meta information fields are `author`, `rights`, or `publisher`, and typical content information fields are `title`, `description`, `subject`, or `alternative`. In our experiments we focus on the content information fields. A major difficulty is the sparsity of the available information: for many records only few fields are given.

The user's information need is specified by 50 topics that are provided by CLEF in the three main languages of the target collections, namely English, German, and French. A topic consists of two fields: a `title`, containing 2-4 keywords, and a `description`, containing 1-2 sentences that specify the item of interest in greater detail. The topics are used to construct the queries.

The TEL@CLEF task is divided into a monolingual and a bilingual subtask. The aim in both subtasks is to retrieve documents (library catalog records) from the target collections, which are most relevant to a query; for each query the results are submitted as a ranked list of documents. In the monolingual subtask the language of the query and the main language of the collection are the same, while in the bilingual subtask

Table 1. Statistics of the three target collections used in the TEL@CLEF task: British Library, BL; Österreichische Nationalbibliothek, ONB; and Bibliothèque nationale de France, BNF.

	BL	**ONB**	**BNF**
main language	English	German	French
# documents	1 000 100	869 353	1 000 100
# documents with `title`	1 000 042	829 675	1 000 095
average length of `title` per document	8.033	5.500	17.124
# documents with `description`	518 493	0	1 000 100
average length of `description` per document	6.222	0	10.095
# documents with `subject`	671 544	602 580	368 788
average length of `subject` per document	7.032	8.373	10.833
# documents with `alternative`	78 679	404 415	0
average length of `alternative` per document	5.491	8.158	0
# documents without content information	20	37 564	0

the language of the query is different from the main language of the collection. We submitted runs for both subtasks and for all three languages.

3 Cross-Language Explicit Semantic Analysis

Cross-Language Explicit Semantic Analysis, CL-ESA, is a generalization of the Explicit Semantic Analysis, ESA [2], and was proposed by Potthast et al. [3]. This section presents a formal definition of the CL-ESA model that reveals its close connection to the Generalized Vector Space Model, GVSM [5]: the ESA model and the GVSM can be transformed into each other [1]. It follows immediately that this is also true for the CL-ESA model and the cross-lingual extension of the Generalized Vector Space Model, CL-GVSM [6].

3.1 Formal Definition

Let d_i be a real-world document written in language L_i, and let $\mathbf{d}_i$ be a bag-of-word-based representation of d_i, encoded as a vector of normalized term frequency weights over a universal term vocabulary V_i. V_i contains all used terms for language L_i. A set $\mathbf{D}_i$ of document representations defines a term-document matrix A_{D_i}, where each column in A_{D_i} corresponds to a vector $\mathbf{d}_i \in \mathbf{D}_i$.

Definition 1 (ESA Representation [1]). *Let D_i^* be a collection of index documents written in language L_i. The ESA representation $\mathbf{d}_{iESA}$ of a document d_i with representation $\mathbf{d}_i$ is defined as follows:*

$$\mathbf{d}_{iESA} = A_{D_i^*}^T \cdot \mathbf{d}_i, \tag{1}$$

where A^T designates the matrix transpose of A.

The rationale of this definition becomes clear if one considers that the weight vectors $\mathbf{d}_i^* \in \mathbf{D}_i^*$ and $\mathbf{d}_i$ are normalized: $||\mathbf{d}_i^*|| = ||\mathbf{d}_i|| = 1$, for each $\mathbf{d}_i^* \in \mathbf{D}_i^*$. Hence, each entry in the ESA representation $\mathbf{d}_{iESA}$ of a document d_i is the cosine similarity between $\mathbf{d}_i$ and some vector $\mathbf{d}_i^* \in \mathbf{D}_i^*$. Put another way, d_i is compared to each index document in D_i^*, and $\mathbf{d}_{iESA}$ is comprised of the respective cosine similarities.

Definition 2 (CL-ESA Similarity). *Let $\mathcal{L} = \{L_1, \ldots, L_k\}$ denote a set of natural languages, and let $\mathcal{D}^* = \{D_1^*, \ldots, D_k^*\}$ be a set of index collections where each $D_i^* \in \mathcal{D}^*$ is a list of index documents written in language $L_i \in \mathcal{L}$. $\mathcal{D}^*$ is a document-aligned comparable corpus, i.e., for each language $L_i \in \mathcal{L}$ the n-th index document in $D_i^* \in \mathcal{D}^*$ describes the same concept. The CL-ESA similarity, $\varphi_{CL-ESA}(q_j, d_i)$, between a query q_j in language L_j and a document d_i in language L_i is computed as cosine similarity φ of the ESA representations of q_j and d_i:*

$$\varphi_{CL-ESA}(q_j, d_i) = \varphi(\mathbf{q}_{j\,ESA}, \mathbf{d}_{i\,ESA}) = \varphi(A_{D_j^*}^T \cdot \mathbf{q}_j, A_{D_i^*}^T \cdot \mathbf{d}_i) \tag{2}$$

Due to the alignment of the index collections D_j^* and D_i^* the ESA representations of q_j and d_i are comparable. Definition 2 is equivalent to the definition of the CL-GSVM similarity $\varphi_{CL-GVSM}(q_j, d_i)$ given in [6], which means that, in analogy to [1], the CL-ESA model and the CL-GVSM can be directly transformed into each other:

$$\varphi_{CL-ESA}(q_j, d_i) = \varphi(A_{D_j^*}^T \cdot \mathbf{q}_j, A_{D_i^*}^T \cdot \mathbf{d}_i) = \varphi_{CL-GVSM}(q_j, d_i) \tag{3}$$

3.2 Alternative Interpretation

The original idea of the CL-ESA model is to map both query and documents into a multilingual concept space, as it is expressed in Equation 2. Note that Equation 2 can be rearranged as follows:

$$\varphi_{CL-ESA}(q_j, d_i) = \varphi(A_{D_j^*}^T \cdot \mathbf{q}_j, A_{D_i^*}^T \cdot \mathbf{d}_i) = \mathbf{q}_j^T \cdot A_{D_j^*} \cdot A_{D_i^*}^T \cdot \mathbf{d}_i \tag{4}$$

In particular, the matrix $A_{D_j^*} \cdot A_{D_i^*}^T = G_{j,i}$ can be computed in advance since it is independent from a particular q_j or d_i. Hence:

$$\varphi_{CL-ESA}(q_j, d_i) = \mathbf{q}_j^T \cdot G_{j,i} \cdot \mathbf{d}_i \tag{5}$$

The rationale of Equation 5 becomes apparent if one recognizes $G_{j,i} = A_{D_j^*} \cdot A_{D_i^*}^T$ as $|V_j| \times |V_i|$ term co-occurrence matrix. The n-th row in $A_{D_j^*}$ corresponds to the distribution of the n-th term $t_n \in V_j$ over the index documents in D_j^*; likewise, the m-th row in $A_{D_i^*}$ corresponds to the distribution of the m-th term $t_m \in V_i$ over the index documents in D_i^*. Recall that the index documents in D_j^* and D_i^* are aligned. I.e., the value in the n-th row and the m-th column of $G_{j,i}$ quantifies the similarity between the distributions of t_j and t_i given the concepts described by the index documents in D_j^* and D_i^*.

The CL-ESA similarity computation of Equation 5 can be viewed in two ways:

(*i*) As a translation of the query representation $\mathbf{q}_j$ into the space of the document representation $\mathbf{d}_i$: $\varphi_{CL-ESA}(q_j, d_i) = (\mathbf{q}_j^T \cdot G_{j,i}) \cdot \mathbf{d}_i$, or,

(*ii*) as a translation of the document representation $\mathbf{d}_i$ into the space of the query representation $\mathbf{q}_j$: $\varphi_{CL-ESA}(q_j, d_i) = \mathbf{q}_j^T \cdot (G_{j,i} \cdot \mathbf{d}_i)$.

These views are different from the original idea of the CL-ESA model where both the query representation and the document representation are mapped into a common multilingual concept space (see Equation 2). From a mathematical standpoint Equation 2 and Equation 5 are equivalent; however, implementing CL-ESA based on the alternative interpretation yields a considerable runtime improvement in practical retrieval

Table 2. The different interpretations of the CL-ESA model.

	Original interpretation	**Alternative interpretation**	
		View (*i*)	View (*ii*)
$\varphi_{CL-ESA}(q_j, d_i) =$	$\varphi(A^T_{D^*_j} \cdot \mathbf{q}_j, A^T_{D^*_i} \cdot \mathbf{d}_i)$	$(\mathbf{q}_j^T \cdot G_{j,i}) \cdot \mathbf{d}_i$	$\mathbf{q}_j^T \cdot (G_{j,i} \cdot \mathbf{d}_i)$
Runtime complexity	$O(l \cdot \lvert D^* \rvert + \lvert D^* \rvert)$	$O(l \cdot \lvert V_j \rvert + l)$	$O(l)$

applications. Table 2 contrasts the interpretations and the related runtime complexities. Here, we assume a closed retrieval situation where from a given target collection D_i in language L_i the most similar documents to a query q_j in language L_j are desired. CLIR with CL-ESA is straightforward: computation of $\varphi_{CL-ESA}(q_j, d_i)$ for each $d_i \in D_i$ and ranking by decreasing CL-ESA similarity.

Under the original interpretation the ESA representations $\mathbf{d}_{i\,ESA}$ of the documents $d_i \in D_i$ can be computed in advance. At retrieval time the query is mapped into the concept space in $O(l \cdot |D^*|)$, where l denotes the number of query terms. The computation of the cosine similarity between the ESA representations $\mathbf{q}_{j\,ESA}$ and $\mathbf{d}_{i\,ESA}$ requires $O(|D^*|)$. Under the alternative interpretation the matrix $G_{j,i}$ can be computed in advance. Note that in practical applications $l \ll |D^*|$, since a reasonable index collection size $|D^*|$ is 10 000, which shows the substantial performance improvement under the alternative interpretation and View (*ii*) .

3.3 Usage in TEL@CLEF

In this subsection we describe implementation details of the CL-ESA model we used in our submission. The following parameter setting was determined by analyzing unofficial experiments of the TEL@CLEF 2008 dataset.

Query and Document Construction. We use the original words of both topic fields, `title` and `description`, as queries. The documents are constructed by merging the text of the three record fields `title`, `subject`, and `alternative`. We assume that the language of these fields is the same within one record; however, this assumption may be violated in some cases since the collections contain multilingual records. Records containing non of these fields are omitted in the experiments (see Table 1).

Index Collection. As index collection Wikipedia is employed. We restrict the multilinguality of our model to the three main languages of the target collections: English, German, and French. Based on a Wikipedia snapshot from March 2009 about 169 000 articles per language can be aligned and fulfill several filter criteria, e.g., to contain more than 100 words or not to be a disambiguation or redirection page. All articles are used as index documents. As term weighting schema $tf \cdot idf$ is used. Query and document words are stemmed using the Snowball stemmers. To speed-up the CL-ESA similarity computation all values below a threshold of $\epsilon = 0.025$ are discarded.

Language Detection. While the language of the queries is determined by the corresponding topics the language of the documents is unknown since the collections are multilingual and no language meta information is provided. In the experiments we resort to a simple "detection by stop words" approach that employs a stop word list for

each of the three main languages and counts for each list the occurrences of the particular stop words within a document. A document is expected to have the language of the list with the highest count; if the detection is inconclusive the main language of the collection is assumed.

4 Cross Querying

Cross querying is a straightforward approach for CLIR systems. We subsume the fields of a topic in one query which is translated in the other languages. With each of the translations we compute a set of rankings by retrieving against each document field. The rankings are merged with respect to their cosine similarities. Additionally, the scores are multiplied by a boosting constant.

Definition 3 (Cross Querying). *Let $\mathcal{L} = \{L_1, \ldots, L_k\}$ denote a set of natural languages and let $\mathcal{F} = \{F_1, \ldots, F_k\}$ denote a set of document fields. $lang : D \to \mathcal{L}$, $lang(d) \mapsto L_i$ estimates the language of a document d. $\mathbf{d}, \mathbf{q}$, and $\mathbf{q}_{L_i}$ are the representations of a document d, a query q and the translation of q in language L_i. Then the cross querying similarity, $\varphi_{CQ}(q, d)$, of a query q and a document d is defined as follows:*

$$\varphi_{CQ}(q,d) = \sum_{F_i \in \mathcal{F}} \Big(b \cdot \varphi(\mathbf{q}_{lang(d)}, \mathbf{d}_{F_i}) \; + \sum_{\substack{L_i \in \mathcal{L}, \\ L_i \neq lang(d)}} \varphi(\mathbf{q}_{L_i}, \mathbf{d}_{F_i}) \Big), \tag{6}$$

where φ is the cosine similarity and b the boosting constant.

The name "Cross Querying" reflects the fact that $|\mathcal{L}| \times |\mathcal{F}|$ rankings are merged by querying in each language in each field. The applied parameters are as follows:

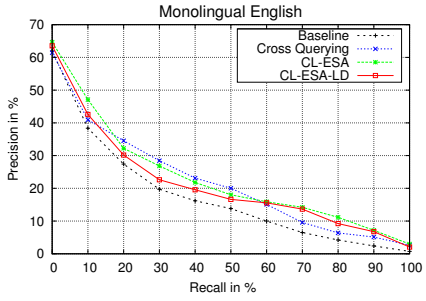
Query and Document Construction. The words of both topic fields, `title` and `description`, are used as queries and translated to each $L_i \in \mathcal{L}$, with $\mathcal{L} = \{German, French, English\}$. The selection of the document fields corresponds to `title` and `subject`. As term weighting schema $tf \cdot idf$ is used. Query and document words are stemmed using the Snowball stemmers while stop words are removed. The queries are translated with Google Translate; the boosting constant b is based on the unofficial evaluation on the TEL@CLEF 2008 dataset.

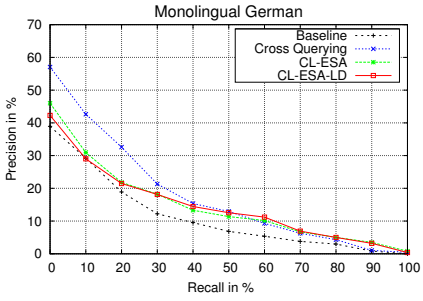
Language Detection. In order to estimate the language of d with $lang(d)$ we take the corpus language of the associated evaluation run.

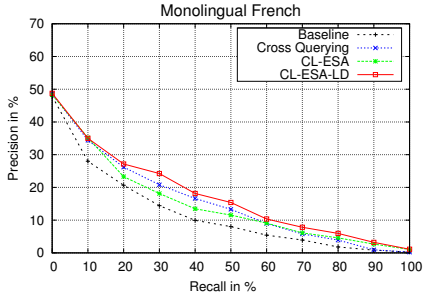
5 Evaluation Results

The results of the monolingual subtask and the bilingual subtask are shown in Figure 1 and Figure 2 respectively.

We submitted an additional baseline to the monolingual subtask using state-of-the-art retrieval technology: since in this subtask the language of the topics is equal to the main language of the target collection, the ranking is based on the cosine similarities of the $tf \cdot idf$-weighted bag-of-words representations of the topics and the documents.

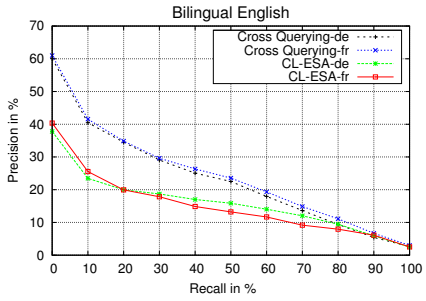


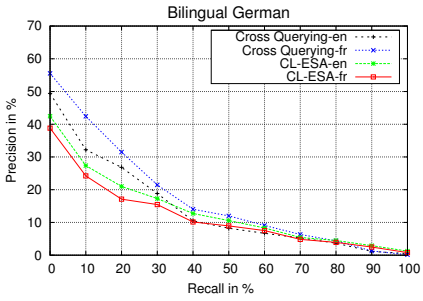


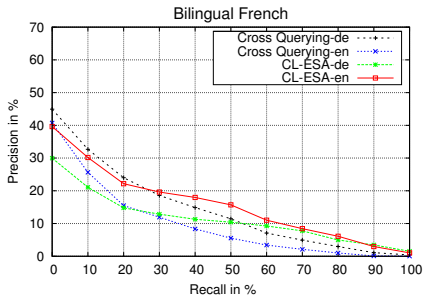


	English	German	French
Baseline	0.158	0.100	0.110
Cross Querying	0.200	**0.164**	0.145
CL-ESA	**0.215**	0.137	0.142
CL-ESA-LD	0.195	0.134	**0.163**

Figure 1. Evaluation results of the monolingual runs. The plots show the standard recall levels vs. interpolated precision. The table show the results in terms of mean average precision, MAP.







	English	German	French
Cross Querying-en	-	0.129	0.132
Cross Querying-de	0.215	-	0.087
Cross Querying-fr	**0.225**	**0.158**	-
CL-ESA-en	-	0.124	**0.145**
CL-ESA-de	0.144	-	0.104
CL-ESA-fr	0.139	0.108	-

Figure 2. Evaluation results of the bilingual runs. The plots show the standard recall levels vs. interpolated precision. The table show the results in terms of mean average precision, MAP.

Each plot in Figure 1 corresponds to one target collection and shows the baseline along with the results achieved under Cross Querying, CL-ESA, and CL-ESA with automatic language detection, CL-ESA-LD. Both Cross Querying and CL-ESA gain a higher MAP than the baseline. The variation between the two approaches is small, except for the German collection where Cross Querying outperforms CL-ESA at low recall levels. At higher recall levels CL-ESA is better, which explains a slightly higher MAP on the English and the French collections. Using CL-ESA along with the automatic language detection improves the performance only for the French collection, which indicates that this collection contains a larger fraction of non-French documents.

In the bilingual subtask the language of the queries is different from the main language of the target collection. Each plot in Figure 2 corresponds to one target collection that is queried in the two other languages, using both Cross Querying and CL-ESA. For example, in the plot "Bilingual English" the graph for "CL-ESA-de" shows the results of querying the English collection with German topics using the CL-ESA. Cross Querying achieves nearly the same or even higher results compared to the monolingual situation, whereas the CL-ESA performs worse in contrast to the monolingual results.

6 Conclusion and Future Work

The evaluation results for the TEL@CLEF task show that both CLIR approaches CL-ESA and Cross Querying are able to outperform the monolingual baseline—though the absolute results are still improvable. Furthermore, we have presented a formal definition and an alternative interpretation for the CL-ESA model, which is interesting for real-world retrieval applications since it reveals how the computational effort for CL-ESA can be shifted from the query phase to a preprocessing phase.

As for future work, CL-ESA and Cross Querying will benefit if more languages are taken into account. Currently, German, English, and French are used, but the target collections comprise more languages. For documents from other languages an inconsistent CL-ESA representation is computed. CL-ESA also needs a reliable language detection mechanism in order to compute a consistent representation; note that we used a rather simple approach in our experiments.

References

1. Maik Anderka and Benno Stein. The ESA Retrieval Model Revisited. In *Proc. of SIGIR 2009*, pages 670–671.
2. Evgeniy Gabrilovich and Shaul Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proc. of IJCAI 2007*, pages 1606–1611.
3. Martin Potthast, Benno Stein, and Maik Anderka. A Wikipedia-Based Multilingual Retrieval Model. In *Proc. of ECIR 2008*, pages 522–530.
4. Philipp Sorg and Philipp Cimiano. Cross-lingual Information Retrieval with Explicit Semantic Analysis. In *Working Notes for the CLEF 2008 Workshop*.
5. S. K. M. Wong, Wojciech Ziarko, and Patrick C. N. Wong. Generalized Vector Spaces Model in Information Retrieval. In *Proc. of SIGIR 1985*, pages 18–25.
6. Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, and Robert E. Frederking. Translingual Information Retrieval: Learning from Bilingual Corpora. *Artif. Intell.*, 103(1-2):323–345, 1998.