Explicit Versus Latent Concept Models for Cross-Language Information Retrieval

Philipp Cimiano
WIS
TU Delft
p.cimiano@tudelft.nl

Antje Schultz, Sergej Sizov
ISWeb
Univ. Koblenz-Landau
antjeschultz@uni-koblenz.de
sizov@uni-koblenz.de

Philipp Sorg
AIFB
Univ. Karlsruhe
sorg@kit.edu

Steffen Staab
ISWeb
Univ. Koblenz-Landau
staab@uni-koblenz.de

Abstract

The field of information retrieval and text manipulation (classification, clustering) still strives for models allowing semantic information to be folded in to improve performance with respect to standard bag-of-word based models. Many approaches aim at a concept-based retrieval, but differ in the nature of the concepts, which range from linguistic concepts as defined in lexical resources such as WordNet, latent topics derived from the data itself - as in Latent Semantic Indexing (LSI) or (Latent Dirichlet Allocation (LDA) - to Wikipedia articles as proxies for concepts, as in the recently proposed Explicit Semantic Analysis (ESA) model. A crucial question which has not been answered so far is whether models based on explicitly given concepts (as in the ESA model for instance) perform inherently better than retrieval models based on "latent" concepts (as in LSI and/or LDA). In this paper we investigate this question closer in the context of a cross-language setting, which inherently requires concept-based retrieval bridging between different languages. In particular, we compare the recently proposed ESA model with two latent models (LSI and LDA) showing that the former is clearly superior to the both. From a general perspective, our results contribute to clarifying the role of explicit vs. implicitly derived or latent concepts in (cross-language) information retrieval research.

1 Introduction

Text-centered tasks such as document classification, clustering and information retrieval all suffer from the so called *vocabulary mismatch* problem, i.e. the problem that documents might be semantically similar (or a document might be relevant for a query) in spite of the fact that the specific terms used (the vocabulary) differ substantially (see [Furnas *et al.*, 1987] on this issue). As a consequence, as overlap with respect to terms is not a necessary condition for semantic relevance or similarity, in some cases methods relying on the bag-of-words model show a poor performance. In fact, the bag-of words model typically assumed in document classification, document clustering and information retrieval inherently suffers from this problem as dimensions in the vector space are inherently orthogonal.

An extreme case of the *vocabulary mismatch* problem can be found in settings where content needs to be matched across languages. Such settings are found, for example, in the by now well-known cross-language information retrieval (CLIR) task, where queries and documents can be in different languages.

In order to overcome the vocabulary mismatch problem, several solutions have been suggested:

1. **latent model**: trying to overcome the orthogonality of dimensions inherent in the bag-of-words model by computing latent dimensions or "concepts" inherent in the data, thus building meaningful groupings beyond single words. Typically, some form of Singular-Value-Decomposition (SVD) is applied to the original document-term matrix to find such latent concepts. Some approaches also define concepts in a probabilistic fashion, such as in Probabilistic Latent Semantic Indexing (PLSI) [Hofmann, 1999].
2. **explicit model**: indexing texts with respect to externally given (explicit) concepts. Generalizations across words can be captured when words can be mapped to more than one category. A recent very prominent example of such a model is the Explicit Semantic Analysis (ESA) method in which texts are indexed with respect to Wikipedia articles as concepts [Gabrilovich and Markovitch, 2007].
3. **mixed models**: adopting the bag-of-words model but extending the standard bag-of-words vector by additional external categories derived from WordNet or some other thesaurus (see [Hotho *et al.*, 2003]).
4. **relatedness models**: incorporating some notion of semantic relatedness between words into the retrieval process. An example for such technique can be found in the approach of Gurevych et al. [Müller and Gurevych, 2008].

As two instances of the above models, we consider the explicit model instantiated by Explicit Semantic Analysis (ESA) as well as the latent model instantiated by LSI/LDA. ESA is by now a prominent representative of explicit approaches. LSI is a prominent representative of latent concept models based on algebraic matrix transformations (Singular

Value Decomposition of the term-document relationship matrix, in our case). Finally, LDA can be seen as a typical probabilistic approach to latent concept computation (i.e. generative probabilistic language model).

As we are concerned with cross-language IR in this paper, we build on CL-LSI [Dumais *et al.*, 1997] as well as on an extension of ESA to a cross-lingual setting (CL-ESA) as described in [Sorg and Cimiano, 2008]. In essence, CL-LSI assumes parallel texts and the original document-term matrix represents each (parallel) document with terms of the different languages. The process of singular value decomposition then yields latent dimensions encompassing words from different languages. In the case of CL-ESA, documents are indexed with respect to their language's Wikipedia articles as in ESA, but the resulting vectors are mapped to vectors for other Wikipedias relying on Wikipedia's cross-lingual structure linking articles to their corresponding articles across languages.

Our focus in this paper is on the empirical comparison of the three mentioned models (CL-ESA, CL-LSI and CL-LDA) on a cross-language retrieval task. The motivation for this comparison stems from the fact that ESA was identified to be superior with respect to LSI on the task of computing relatedness between words [Gabrilovich and Markovitch, 2007]. Interesting and in our opinion straightforward research questions are whether i) ESA performs better than LSI/LDA also on document retrieval tasks and ii) in particular on cross-language retrieval tasks where LSI for instance has been shown to achieve reasonable results (see [Dumais *et al.*, 1997]).

The remainder of the paper is structured as follows: in Section 2 we describe a generic framework for cross-lingual concept-based retrieval, which is later instantiated for CL-ESA (Section 3) as well as CL-LSI/LDA (Section 4). In Section 5 we describe our data, experimental settings and results. Section 6 discusses related work before concluding.

2 General Framework

The general framework for cross-lingual retrieval we build on in this paper has the following characteristics:

- it is vector based and builds on the cosine similarity to assess the relevance between documents and queries
- documents are represented as vectors in a certain "concept space" which abstracts from the various languages considered

The *concept space* is defined by a set of concepts C which might be explicit and external (as in ESA) or data-derived as in (P-)LSI/LDA. The different frameworks then need to define the following components of the vector-based and concept-based retrieval: i) the concept space C, and ii) the function mapping documents $d \in D$ into the concept space, i.e. define the function $\Phi : D \rightarrow \mathbb{R}^{|C|}$. Here we assume that the resulting vectors are thus real-valued.

Further, our framework distinguishes between two document collections: i) the document collection B providing *background knowledge* used to map documents into the concept space and ii) the actual target document collection D on which the retrieval is performed. In all our instantiations, the Wikipedias corresponding to the languages English, German and French or the target document collection itself will represent the background knowledge collection B. However, our framework is generic enough to accommodate other collections.

In what follows we discuss the various instantiations of this general framework, discussing in particular the properties of the concept space.

3 Cross-language Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) [Gabrilovich and Markovitch, 2007] attempts to index or classify a given document d with respect to a set of explicitly given external categories. It is in this sense that ESA is explicit compared to approaches which aim at representing texts with respect to latent topics or concepts, as done in Latent Semantic Indexing (LSI) (see [Deerwester *et al.*, 1990]). Gabrilovich and Markovitch have outlined the general theory behind ESA and in particular described its instantiation to the case of using Wikipedia articles as external categories. We will build on this instantiation as described in [Gabrilovich and Markovitch, 2007] which we briefly summarize in the following.

Explicit Semantic Analysis takes as input a document d and maps it to a high-dimensional real-valued vector space. This vector space is spanned by a Wikipedia database $W_k = \{a_1, \ldots, a_n\}$ in language L_k such that each dimension corresponds to an article a_i. ESA is explicit in the sense that the concept space C corresponds exactly to the article space of Wikipedia, i.e. $C := W_k$.

This mapping is given by the following function: $\Phi_k^{ESA} : D \rightarrow \mathbb{R}^{|W_k|}$, where $\Phi_k^{ESA}(d) := \langle as^{ESA}(d, a_1), \ldots, as^{ESA}(d, a_n) \rangle$. The value $as(d, a_i)$ in the ESA vector of d expresses the *strength of association* between d and the Wikipedia article a_i.

One approach to define such an association strength function as is to use a TF.IDF function based on the Bag-of-Words (BOW) model of the Wikipedia articles [Gabrilovich and Markovitch, 2007]. The association strength can then be computed as the sum of the TF.IDF values of the article a_i for all words of $d = \langle w_1, \ldots, w_s \rangle$: $as^{ESA}(d, a_i) := \sum_{w_j \in d} TF.IDF_{a_i}(w_j)$

The ESA framework has been extended in [Sorg and Cimiano, 2008] and [Potthast *et al.*, 2008] for the case of cross-language retrieval by mapping the vector representation of documents from one Wikipedia article space to a Wikipedia from another language. This is achieved by exploiting Wikipedia's language links. A function $m_{i \rightarrow j} : W_i \rightarrow W_j$ is assumed to return for an article in Wikipedia i the corresponding article in Wikipedia for language j[1].

In fact, given a text $d \in D$ in language L_i, it turns out that we can simply index this document with respect to any of the other languages $L_1, .., L_n$ we consider by transforming the vector $\Phi_i^{ESA}(d)$ into a corresponding vector in the vector

[1] While this is not the case in general, in our specific settings (see Section 5) $m_{i \rightarrow j}$ can indeed be assumed to be a bijective function.

space that is spanned by the articles of Wikipedia in the target language. Thus, given that we consider n languages, we have n^2 mapping functions of the type: $\Psi_{i\to j} : \mathbb{R}^{|W_i|} \to \mathbb{R}^{|W_j|}$

$$\Psi_{i\to j}(\langle v_1, \ldots, v_{|W_i|}\rangle) := \langle v_{m_{i\to j}{}^{-1}(1)}, \ldots, v_{m_{i\to j}{}^{-1}(|W_j|)}\rangle$$

In order to speed up processing and yield more compact vectors, we consider only the top n dimensions of the ESA vectors by using projections Π_n of our vectors selecting only the n dimensions with the highest values. The parameter n will be varied in our experiments.

Given the above settings, it should be straightforward to see how the actual retrieval works. The cosine between a query q_i in language L_i and a document d_j in language L_j is calculated as: $cos(q_i, d_j) := cos(\Pi_n(\Psi_{i\to j}\Phi_i(q_i)), \Pi_n((\Phi_j(d_j))))$. We used our own ESA implementation as used already in [Sorg and Cimiano, 2008].

4 Latent Semantic Analysis for CLIR

4.1 Latent Semantic Indexing (LSI)

LSI is a well known approach for extracting concepts from a given text corpus [Deerwester *et al.*, 1990]. It is based on singular value decomposition (SVD), a technique from Linear Algebra. As a full SVD is a loss-free decomposition of a matrix M, which is decomposed into two orthogonal matrices U and V (left and right singular vectors) and a diagonal matrix Δ (singular values) estimating less singular values and their corresponding singular vectors leads to an approximation of M by: $M \approx \tilde{U} * \Delta * \tilde{V}^T$. The set of concepts C is given implicitly by the columns of U, i.e. $C = col(U)$. U covers the term-concept-space by holding a weight for the correlation of each term-concept-pair. Analogously, the document-concept-space V contains a weight for the document-concept-correlation. Since the dimension of Δ corresponds to the number of concepts and singular values are derived in descending order a reduced SVD results in the most relevant concepts.

In the term-document-matrix, each document of the background knowledge collection B is represented as a term-vector using $TF.IDF$-values as weights. By reducing the dimension of the model, LSI brings related terms together and forms concepts. In this new space documents are no longer represented by terms but by concepts. New documents (e.g. from the retrieval document collection) or queries are represented in terms of concepts by "folding them in into the LSI model" by multiplying their term-vector with U. The document-concept mapping is thus defined by the following function: $\Phi(d) = U^T * d$. In contrast to ESA, documents are mapped to a representation of lower dimension. Similarity between documents or query and documents is computed via the cosine-measure in our LSI implementation.

For Cross Language LSI (CL-LSI), a parallel text corpus consisting of parallel (similar) documents in different languages is used, e.g. Wikipedia as cross-lingual corpus or parallel corpora of automatically or manually translated documents. Each document in this corpus is constructed as a multiset of words by merging the words for the same document in different languages. Such a multiset is treated as one document in the document-term-matrix consisting of several languages, from which a multi-lingual feature space is built [Dumais *et al.*, 1997] and to which standard singular value decomposition can be applied, yielding multilingual concepts.

4.2 Latent Dirichlet Allocation

As an instance of the probabilistic latent topic models, we consider the LDA based generative model. The basic idea of this approach is to abstract from particular words and to represent multi-lingual documents by mixtures over a set C of latent concepts $c_1 \ldots c_k$ (i.e. hidden document-specific themes of interest), whereby each concept is characterized by a fixed conditional distribution over words from different languages.

LDA assumes that all multi-lingual terms (both observed and previously unseen) are generated by randomly chosen latent concepts. In contrast to the SVD used in LSI, LDA has a well founded probabilistic background and tends to result in more flexible model fitting [Blei *et al.*, 2003]. It allows resources to belong to multiple latent concepts with different degrees of confidence and offers a natural way of assigning probabilistic feature vectors to previously unseen resources.

In line with [Blei *et al.*, 2003], the content of the particular document is generated by selecting a multinomial distribution over concepts given the Dirichlet prior. For each term, a concept is generated from the document-specific concept distribution, and then a keyword is generated from the discrete distribution for that concept as follows:

1. The number of words in the multi-lingual document is chosen: $n \sim Poisson(\xi)$
2. The keyword generating parameter is chosen: $\theta \sim Dir(\alpha)$
3. For each of the document terms $t_i, i = 1 \ldots n$:
   - The generative concept for t_i is chosen: $c_i \sim Multinomial(\theta)$.
   - The term t_i is generated using a multinomial probability with parameter β conditioned on c_i: $p(t_i|c_i, \beta)$

For the multi-lingual retrieval scenario, the LDA approach can be instantiated as follows. In the first step, the background knowledge collection B is constructed as in the LSI case (Section 4.1). The resulting 'mixed' multisets are considered as 'training documents' and used for fitting the corpus-level properties α and β which are estimated using the variational Expectation Maximization (EM) procedure [Blei *et al.*, 2003]. In the process of corpus analysis we also obtain the document-level variables θ, sampled once per multi-lingual document. As a result we obtain the posterior distribution of the hidden concepts $c_1 \ldots c_k$ given a document d:

$$p_d(\theta, \vec{c}|\vec{t}, \alpha, \beta) = \frac{p_u(\theta, \vec{c}, \vec{t}|\alpha, \beta)}{p_u(\vec{t}|\alpha, \beta)} \quad (1)$$

Given the estimated model parameters, the similarity between new documents in different languages can be computed. First, each document is mapped to a concept-based representation by Φ. In the LDA approach Φ estimates the distribution of the hidden concepts $c_1 \ldots c_k$ (1) using the variational EM

procedure [Blei *et al.*, 2003]. The estimated probabilities are considered as characteristic features of the document in the latent concept-based feature space: $\Phi(d) = p_d(\theta, \vec{c}|\vec{t}, \alpha, \beta)$. Our expectation is that similar (or identical) documents in different languages show the similar behavior in terms of probabilities over latent concepts (1). We notice that the cosine similarity measure enforces in this case the intuitively expected IR-like similarity estimation behavior, whereby the similarity of the document to itself is maximized and the orthogonal document feature vectors (i.e. document vectors with disjoint sets of non-zero topic probabilities) have zero similarity.

5 Experiments

The key objective of our experimental evaluation was to compare the appropriateness of discussed latent and explicit topical models for cross-lingual search and retrieval scenarios. In doing so, we compared different representation models using the same utility (ranking) function for retrieval of similar multi-lingual documents. In our comparative experiments we used our own implementation of the ESA framework, the Octave SVD implementation [2], and the LingPipe LDA implementation [3].

5.1 Methodology

For the comparison of the presented CLIR methods we used a standard mate retrieval setup. For mate retrieval evaluation, a text corpus with parallel documents in different languages is used. Taking the document in one language as a query, similar documents in another language are retrieved. It is assumed that the translated version (mate) is most similar to the document itself and therefore should appear on top of the ranked result list. Consequently, the observed position of the mate can be used as a comparison yardstick. Since document mates are explicitly known, no manual relevance assessment is needed for this kind of evaluation. In all evaluations, we used the cosine similarity measure for estimating similarity between feature vectors of the query document and of test documents.

As quality criteria, we considered top-1 and top-10 accuracy as well as the Mean Reciprocal Rank (MRR). Top-k accuracy is the fraction of query documents, for which the mate document was placed among the first k matches in the ranked list. MRR estimates the average position of the mate document, with higher values corresponding to better retrieval results.

5.2 Datasets

In our experiments we used two multi-lingual reference test collections D, Multext and JRC-Acquis. The Multext collection is a parallel corpus derived from the Multext project[4]. It contains ca. 3100 question/answer pairs from the Official Journal of European Community (JOC), translated into five languages. The JRC-Acquis [5] multilingual parallel corpus is

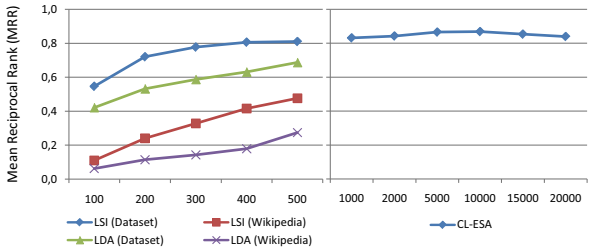
[2] http://www.gnu.org/software/octave/

[3] http://alias-i.com/lingpipe/

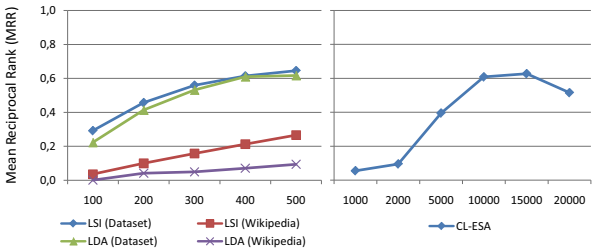
[4] http://aune.lpl.univ-aix.fr/projects/MULTEXT/

[5] http://langtech.jrc.it/JRC-Acquis.html

Figure 1: Results for the mate retrieval experiments on English and French documents for different topic numbers for LSI/LDA and different ESA vector lengths for CL-ESA.



(a) Multext dataset



(b) JRC-Acquis dataset

a collection of approx. 21,000 legislative documents of the European Union in 22 European languages.

For the Multext corpus, we used all documents available in three target languages English, German, and French (3055). Each document in one language was used as a query for searching in all documents of another language. For the JRC-Acquis corpus we randomly selected 3000 multi-lingual documents, fairly comparable to the entire size of the Multext corpus, and performed the similar evaluation on this fixed subset, but searching on the complete set of 21,000 documents, which explains why the results on Multext are better than those on JRC-Acquis.

For constructing the background document collection B, we used the publicly available English, German and French snapshots of Wikipedia [6]. We analyzed cross-language links between Wikipedia articles and restrictively used only articles correctly linked to each other across all three languages. Using the snapshot by 03/12/2008 for English, 06/25/2008 for French, and 06/29/2008 for German, we obtained the aligned collection of 166,484 articles in all three languages.

All mentioned collections were prepared using common IR-like preprocessing steps including text extraction, elimination of stopwords, special characters and extremely short terms (length < 3), and stemming using language-specific Snowball stemmers[7]. For the corresponding German collections, we additionally applied a dictionary-based compound splitter based on [Koehn and Knight, 2003], using the entire German Wikipedia corpus as a reference dictionary.

Even after the preprocessing the Wikipedia corpus was too huge and too sparse for an efficient LSI. After claiming that each document and it's mate should have between 50 and

[6] http://download.wikimedia.org/backup-index.html

[7] http://snowball.tartarus.org

500 terms we skipped all terms which appear less than 15 times. Finally we used all documents containing at least 75 terms, e.g. the number of English-French document pairs was reduced to 54764. These restrictions were applied to each language pair separately. Because of sparseness of the term-document-matrices a similar process was applied to the JRC dataset.

5.3 Results

Both of the test collections, Multext and JRC-Acquis, are parallel corpora with explicitly aligned document versions in different languages, and are therefore directly applicable to the mate retrieval setting. For every CLIR method, we considered English, German and French as baseline languages and performed accordingly 6 series of experiments with all possible language pairs. The results for one language pair in both directions (e.g. English-German and German-English) were then averaged.

Figure 1 shows sample MRR results for mate retrieval experiments between English and French documents. ESA reaches its peak performance at about 10,000 dimensions considered, which shows that generally it needs a minimum number of "concepts" to perform reasonably, but clearly also reaches a point were further concepts start introducing noise somewhere after 10,000 documents. For latent topic models, the accuracy tends to increase from 100 to 500 topics. The exploration of computationally much more expensive latent models ends with 500 topics, due to computational limitations of our servers. In the experiments we report below, we used these optimal settings (10,000 for ESA, 500 for latent models) for all language pairs.

Table 1 shows the corresponding summary of achieved results. It includes the Top-1 and Top-10 accuracy as well as the Mean Reciprocal Rank of the mate. In addition, for latent topic methods we compared the retrieval characteristics for different choices of the background knowledge collections B, namely Wikipedia vs. the test collection itself at the ratio of 60% for learning and 40% for testing. The results show that w.r.t. the Top-1 accuracy measure, ESA outperforms both other models on all language pairs of the Multext dataset and on the en-fr and en-de pairs of the JRC-Acquis dataset, but not on de-fr. With respect to Top-10, LSI is in all cases better than ESA, but only when it has been trained on the retrieval document collection itself. When LSI is trained on Wikipedia as an aligned corpus, results are in all cases worse. This shows that ESA is indeed superior as it does not rely on an aligned corpus to be trained on to deliver good results.

5.4 Discussion

Our results show that the use of Wikipedia as a background knowledge source B leads to significantly worse results for latent topic models (in contrast to the case when they are trained on the retrieval document collection). The explanation of this phenomenon is twofold. On one hand, Wikipedia is not a fully parallel corpus and linked articles may show substantial variation in size, quality, and vocabulary. On the other hand, there is a serious vocabulary mismatch between Wikipedia and our thematically focused test collections. For instance, in the Multext collection, 4713 English terms (44%), 8055 German terms (53.8%) and 7085 French terms (53.6%) are not covered by Wikipedia articles at all. We also assume that the performance of LDA observed in our experiments can be further improved by heuristic model tuning, including optimization of concentration parameters for Dirichlet priors, or smoothing of estimated multinomial parameters (as described in [Blei *et al.*, 2003]).

Overall, it can be claimed that ESA clearly outperforms LSI / LDA unless the latter are trained on the document collection and not on Wikipedia. The availability of aligned corpora is a serious restriction, so that ESA is clearly the preferred model here as it delivers reasonable results requiring no data aligned across languages besides Wikipedia. A further crucial advantage is its excellent scalability: ESA does not require comprehensive computations with nonlinear space/time behavior and can be practically performed within 20-30 min for any desired number of topis in each of our test collections. In contrast, the computation of LSI and LDA models (with significantly lower dimensionality) took between 3h and 7 days.

6 Related Work

Among the most prominent approaches to cross-language information retrieval are translation-based techniques on the one hand and concept-based on the other. Translation-based techniques come in two different modes: either the query is translated into the target language typically by using bilingual dictionaries (see [Levow *et al.*, 2005]) or the documents are translated into the query language by using some fully-fledged translation system. Such systems have been recently shown to be very successful in CLEF campaign evaluations (see [Kürsten *et al.*, 2008]).

The second class of approaches are concept-based as motivated in Section 1. Dumais et al. [1997] also used LSI to compute "latent concepts" inherent in the data collection to perform cross-lingual retrieval similar to the approach in this paper. On a similar mate retrieval task as considered in this paper, a LSI-based system as described in [Littman *et al.*, 1998] achieves a Top-1 accuracy of above 98.3% (English-French) and 98.5% (French-English) using a 982-dimensional representation of the multilingual document space. These higher results are due to the fact that their LSI-based approach is trained on the very same collection where it is evaluated. Recently, several approaches have emerged which rely on explicitly defined concepts for indexing and retrieval. Many of these approaches are based on the seminal work on Explicit Semantic Analysis of Gabrilovich and Markovitch [2007]. They showed that ESA outperforms a bag-of-words baseline as well as LSI on the task of computing semantic relatedness of words. To our knowledge, this has not been shown in the context of a cross-lingual retrieval task. Two very similar approaches to ours are the one of Potthast et al. [2008] as well as Müller and Gurevych [2008]. While Potthast et al. apply ESA similarly to a cross-language retrieval task, their implementation of ESA is different to ours as they use a different association function as^{ESA} based on the cosine of term vectors of documents and articles. Potthast et al. achieved results with Top-1 Accuracy up to over

Table 1: Results for the mate retrieval experiments on the Multext and JRC-Acquis dataset using optimal settings for topic numbers for LSI/LDA (500) and ESA vector lengths (10,000). Evaluation measures are Top-1 and Top-10 Accuracy and Mean Reciprocal Rank.

		en-fr			en-de			de-fr		
Dataset	**Method**	TOP-1	TOP-10	MRR	TOP-1	TOP-10	MRR	TOP-1	TOP-10	MRR
Multext	CL-ESA	**.83**	.94	**.87**	**.72**	.90	**.78**	**.64**	.84	.71
	LSI (Dataset)	.71	**.98**	.81	.60	**.96**	.72	.59	**.97**	**.72**
	LSI (Wikipedia)	.36	.70	.48	.13	.39	.21	.13	.41	.22
	LDA (Dataset)	.11	.24	.69	.04	.11	.48	.05	.12	.47
	LDA (Wikipedia)	.01	.04	.27	.01	.02	.16	.01	.03	.14
JRC-Acquis	CL-ESA	**.56**	.70	.61	**.35**	.49	.40	.27	.42	.32
	LSI (Dataset)	.52	**.87**	**.65**	.29	**.80**	**.45**	**.34**	**.78**	**.49**
	LSI (Wikipedia)	.18	.46	.27	.07	.23	.12	.07	.26	.13
	LDA (Dataset)	.08	.14	.62	.12	.09	.36	.04	.15	.38
	LDA (Wikipedia)	.01	.03	.09	.01	.02	.07	.01	.02	.08

90% (for 100,000 ESA dimensions) for selected JRC-Acquis documents and Wikipedia articles.

7 Conclusion

In this paper we have compared retrieval models based on explicit concepts (ESA in particular) with models based on latent concepts (LSI and LDA in particular) on a cross-language mate retrieval task on two datasets (JRC-Acquis and Multext). We have clearly shown that, unless LSI/LDA are trained on the document collection itself (instead of on the background collection, i.e. Wikipedia in our case), ESA is clearly superior to LSI/LDA both in terms of quality of results, but also in terms of computational performance. We are not aware of any previous comparison of ESA and LSI/LDA w.r.t. cross-language information retrieval tasks, so that our work represents an important contribution to the field and provides an important step towards clarifying the role of explicit vs. latent concepts in information retrieval in general. In future work, we intend to investigate the techniques under discussion in this paper on datasets which do not consist of parallel corpora to show that our results hold beyond the specific settings we have considered here.

Acknowledgements

This work was funded by the German Research Foundation (DFG) under the Multipla project (grant 38457858).

References

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Deerwester *et al.*, 1990] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[Dumais *et al.*, 1997] S. Dumais, T. Letsche, M. Littman, and T. Landauer. Automatic Cross-language Retrieval using Latent Semantic Indexing. In *Proceedings of the AAAI Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[Furnas *et al.*, 1987] G.W. Furnas, T.K. Landauer, L.M. Gomez, and S.T. Dumais. The Vocabulary Problem in Human-System Communication. *Communications of the ACM*, 30(1):964–971, 1987.

[Gabrilovich and Markovitch, 2007] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of IJCAI*, pages 1606–1611, 2007.

[Hofmann, 1999] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of SIGIR*, pages 50–57, 1999.

[Hotho *et al.*, 2003] A. Hotho, S. Staab, and G. Stumme. Ontologies Improve Text Document Clustering. In *Proceedings of ICDM*, pages 541–544, 2003.

[Koehn and Knight, 2003] P. Koehn and K. Knight. Empirical Methods for Compound Splitting. In *Proceedings of EACL*, pages 187–194, 2003.

[Kürsten *et al.*, 2008] J. Kürsten, T. Wilhelm, and M. Eibl. CLEF 2008 Ad-Hoc Track: On-line Processing Experiments with Xtrieval. In *Working Notes of the Annual CLEF Meeting*, 2008.

[Levow *et al.*, 2005] G.-A. Levow, D.W. Oard, and P. Resnik. Dictionary-based Techniques for Cross-language Information Retrieval. *Information Processing and Management*, 41(3):523–547, 2005.

[Littman *et al.*, 1998] M.L. Littman, S.T. Dumais, and T.K. Landauer. *Cross-Language Information Retrieval*, chapter Automatic Cross-Language Information Retrieval using Latext Semantic Indexing, pages 51–62. Kluwer, 1998.

[Müller and Gurevych, 2008] C. Müller and I. Gurevych. Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval. In *Working Notes of the Annual CLEF Meeting*, 2008.

[Potthast *et al.*, 2008] M. Potthast, B. Stein, and M. Anderka. A Wikipedia-Based Multilingual Retrieval Model. In *Proceedings of ECIR*, pages 522–530, 2008.

[Sorg and Cimiano, 2008] P. Sorg and P. Cimiano. Cross-lingual Information Retrieval with Explicit Semantic Analysis. In *Working Notes of the Annual CLEF Meeting*, 2008.