An Experimental Comparison of Explicit Semantic Analysis Implementations for Cross-Language Retrieval

Philipp Sorg[1] and Philipp Cimiano[2]

[1] Institute AIFB, University of Karlsruhe
sorg@kit.edu
[2] Web Information Systems Group, Delft University of Technology
p.cimiano@tudelft.nl

Abstract. Explicit Semantic Analysis (ESA) has been recently proposed as an approach to computing semantic relatedness between words (and indirectly also between texts) and has thus a natural application in information retrieval, showing the potential to alleviate the vocabulary mismatch problem inherent in standard Bag-of-Word models. The ESA model has been also recently extended to cross-lingual retrieval settings, which can be considered as an extreme case of the vocabulary mismatch problem. The ESA approach actually represents a class of approaches and allows for various instantiations. As our first contribution, we generalize ESA in order to clearly show the degrees of freedom it provides. Second, we propose some variants of ESA along different dimensions, testing their impact on performance on a cross-lingual mate retrieval task on two datasets (JRC-ACQUIS and Multext). Our results are interesting as a systematic investigation has been missing so far and the variations between different basic design choices are significant. We also show that the settings adopted in the original ESA implementation are reasonably good, which to our knowledge has not been demonstrated so far, but can still be significantly improved by tuning the right parameters (yielding a relative improvement on a cross-lingual mate retrieval task of between 62% (Multext) and 237% (JRC-ACQUIS) with respect to the original ESA model).

1 Introduction

The quest for a more "semantic" retrieval of information items (documents, videos, music etc.) still represents one of the more challenging research directions in information retrieval today. There have been many approaches so far aiming at incorporating "semantics" into the retrieval process. Prominent examples are those that use thesauri for query expansion. These thesauri can be either manually created as in the case of WordNet [1] or derived from the local document collection (see e.g. [2]). Other approaches integrate semantic relatedness or semantic similarity between words into the retrieval process [3]. Finally, other approaches aim at a concept-based retrieval, where such concepts can be either computed implicitly from the document collection, as in Latent Semantic Indexing [4] or given explicitly by external resources such as WordNet [5].

One very successful approach in the latter direction which has attracted a lot of attention in recent years is the Explicit Semantic Analysis (ESA) model by Gabrilovich

H. Horacek et al. (Eds.): NLDB 2009, LNCS 5723, pp. 36–48, 2010.

and Markovitch [6]. In essence, ESA indexes documents with respect to the Wikipedia article space (as "conceptual" space), indicating how strongly a given word in the document (and by aggregation also the whole document) is associated to a specific Wikipedia article. Gabrilovich and Markovitch instantiate a geometric framework in which each word is represented as a vector of Wikipedia articles and similarity is calculated using the cosine measure, where the $tf.idf$ value of a word in a given Wikipedia article is used as weight of the corresponding dimension in the vector. As a word can be associated to many articles (with different weights), ESA alleviates the vocabulary mismatch problem [7] inherent in the BOW model, where every word corresponds exactly to one dimension, the dimensions being orthogonal. In the ESA model, two words or texts can be semantically related in spite of not having any word in common (but associated to similar Wikipedia articles).

In this paper, we put the ESA model under scrutiny and empirically analyze variants of the original ESA model, both looking at alternatives for calculating the association between Wikipedia articles and words as well as examining alternative retrieval models, in particular based on language modeling approaches as well as probabilistic models. We investigate these variants in the context of a cross-language retrieval task following a cross-lingual extension of ESA (CL-ESA) (see [8] and [9]). We evaluate the ESA variants with respect to the well-known mate retrieval task, i.e. given a parallel corpus, retrieving for each document its parallel document in another language as in [10]. We report experiments on two parallel datasets, the Multext dataset as well as the JRC-ACQUIS corpus on three languages: English, French and German.

Our results show on the one hand that the choice of some parameters (in particular the association strength but also the retrieval model) can have a significant impact and, on the other hand, that, while the settings adopted in the original ESA model are reasonable, its performance can be significantly increased by changing some of the parameters. To our knowledge, there has been no empirical analysis and comparison between different implementation choices before.

The paper is structured as follows: in the following Section 2 we present the ESA model in the standard (monolingual) version (as described in [6]) as well as the cross-lingual formulation along the lines of [8], both for the sake of completeness and to facilitate the understanding of this paper. In Section 3 we then first introduce a generalization of the ESA model which makes explicit the choices that it leaves open and discuss various alternatives for these choices. In Section 4 we then experimentally analyze and present the results of the different variants on a cross-lingual mate retrieval task.

2 Explicit Semantic Analysis (ESA)

2.1 Classical (Monolingual) ESA

Explicit Semantic Analysis (ESA) [6] 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 Analysis (LSA) (see [4,11]). 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 basically build on this instantiation as described in [6], which we briefly summarize in the following.

In essence, 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 $\mathcal{W}_k = \{a_1, \ldots, a_n\}$ in language L_k such that each dimension corresponds to an article a_i. This mapping is given by the following function: $\Phi_k : D \rightarrow \mathbb{R}^{|\mathcal{W}_k|}$ with

$$\Phi_k(d) := \langle as(d, a_1), \ldots, as(d, a_n) \rangle$$

The function as expresses the *association strength* between d and the Wikipedia article a_i. In the original ESA model, as is defined by sum of $tf.idf$ values of all words of $d = \langle w_1, \ldots, w_s \rangle$ in the article a_i multiplied by tf in d:

$$as(d, a_i) := \sum_{w_j \in d} tf_d(w_j) tf.idf_{a_i}(w_j)$$

Essentially the Semantic Interpreter applying ESA described in [6] computes the function Φ. As output we thus get a vector representing the strength of association of a document d with respect to the articles in Wikipedia $\mathcal{W}_k$. These vectors can then be used to assess the similarity between documents at a conceptual level (e.g. using cosine similarity between the articles indexed with respect to the Wikipedia articles, i.e. the vectors yielded by the Φ-function) and have thus a natural application in information retrieval tasks, which we are concerned with in this article.

In the following section, we present the extension to ESA called CL-ESA (Cross-language Explicit Semantic Analysis), which represents a relatively straightforward extension of ESA to a cross-lingual setting presented before [9,8].

2.2 Cross-lingual ESA (CL-ESA)

It has been shown recently that, when instantiating ESA for Wikipedia, one can rely on Wikipedia's language links to transform ESA vectors in one language to another one[1]. This is done by mapping each dimension corresponding to article a in Wikipedia $\mathcal{W}_a$ to the dimension corresponding to article b in Wikipedia $\mathcal{W}_b$ so that there exists a language link from a to b. In the following we therefore assume the existence of a mapping function $m_{a \rightarrow b} : \mathcal{W}_a \rightarrow \mathcal{W}_b$ that maps articles according to the language links to articles in another language. This function is only defined for articles having a language link to the Wikipedia in the target language. To overcome this restriction we will use only that subset of the Wikipedia articles having unique language links to all languages considered, such that the function is actually a bijection. We will describe the Wikipedia subset used in more detail in Section 4.

Given a document $d \in D$ in language L_a, CL-ESA allows to index this document with respect to any of the other languages $L_1, \ldots, L_n$ by transforming the vector $\Phi_a(d) = \langle d_{a_1}, d_{a_2}, \ldots \rangle$ into a corresponding vector in the vector space that

[1] Cross-language links are those that link a certain article to a corresponding article in the Wikipedia database in another language.

is spanned by the Wikipedia articles in the target language. This mapping function $\Psi_{a\rightarrow b} : \mathbb{R}^{|\mathcal{W}_a|} \rightarrow \mathbb{R}^{|\mathcal{W}_b|}$ is calculated as follows:

$$\Psi_{a\rightarrow b}(\Phi_a(d)) := \langle d_{m_{b\rightarrow a}(b_1)}, d_{m_{b\rightarrow a}(b_2)}, \ldots \rangle$$

where b_j are the articles of Wikipedia $\mathcal{W}_b$. This means that $\Psi_{a\rightarrow b}(\Phi_a(d))$ is the ESA representation of d with respect to Wikipedia $\mathcal{W}_b$ based on the ESA representation of d with respect to Wikipedia $\mathcal{W}_a$ and the language links between $\mathcal{W}_a$ and $\mathcal{W}_b$.

Given the above settings, it should be straightforward to see how the actual retrieval works. The cosine between a query q_a in language L_a and a document d_b in language L_b is calculated as:

$$cos(q_a, d_b) := cos(\Psi_{a\rightarrow b}(\Phi_a(q_a)), \Phi_b(d_b))$$

In our settings the query vector is thus mapped to the target language and compared to documents in the target language. This thus gives us an elegant retrieval model which is uniform across languages. A prerequisite for this model is certainly that we know the language of the query and of the different documents in order to know which mapping Ψ should be applied.

3 ESA Variants

We first present the generalization of the ESA model, making the choices for different parameters explicit. This will provide a uniform model to investigate the impact of different parameters on the ESA model. Then, we present the specific alternatives for the different choices that we have experimentally compared in Section 4.

3.1 Generalization

A cross-lingual retrieval model based on ESA can be generalized as follows (q_a is a query and d_b a document in the collection):

$$rel(q_a, d_b) := rel(\Pi(\Psi_{a\rightarrow b}(\Phi_a(q_a))), \Pi(\Phi_b(d_b)))$$

with

$$\Phi(d) := \boldsymbol{d} = \langle as(d, a_1), \ldots, as(d, a_{|\mathcal{W}|}) \rangle$$

The relevant parameters to be instantiated are:

- **Dimension Projection Function** Π: For most implementations of ESA, it is impossible to work with all of the dimensions for which the association strength is greater than 0 (for pragmatic reasons related to efficiency of computation). Therefore, most approaches index a text only with respect to a subset of the relevant dimensions.
- **Association Strength Function** as: The so called association strength function quantifies the degree of association between a document d and a category a_j.
- **Relevance Function / Retrieval Model** rel: Concerning the retrieval model, while the cosine (thus assuming a geometric retrieval model) has been used, other alternatives are possible here.

- **Category System**: ESA relies on the fact that there is some external category system with respect to which words and texts can be indexed. While Wikipedia has been used in most implementations, the originators of ESA have also tested on an alternative category system: the Open Directory Project (ODP)[2], achieving worse results than with Wikipedia. Though the choice of the category system is also crucial, in this work we will rely on the Wikipedia-based implementation as in the context of our cross-lingual retrieval experiments we directly exploit the language links of Wikipedia to map between languages.

This offers a generalized framework for the ESA model allowing different parameters to be explored and to analyze their impact. We will discuss particular implementations of the above functions for which we will also provide experimental evaluation in Section 4.

3.2 Dimension Projection

We will consider the following variants for the dimension projection function Π that have been considered in previous literature (but never been analyzed systematically). As notation we will refer to $\boldsymbol{d}_i$ as the i-th dimension of the ESA vector of d which is the association strength of d to the article a_i. The function α_d defines an order on the indices of the dimensions according to descending values such that $\forall i, j : \; i < j \rightarrow \boldsymbol{d}_{\alpha(i)} \geq \boldsymbol{d}_{\alpha(j)}$, e.g. $\boldsymbol{d}_{\alpha(10)}$ is the 10-th highest value of $\boldsymbol{d}$.

1. **Absolute**, with $\Pi^m_{abs}(\boldsymbol{d})$ being the projected vector by restricting $\boldsymbol{d}$ to the m dimensions with highest values:, i.e. $\alpha(1), \ldots, \alpha(m)$ (as in [9] and [8])
2. **Absolute Threshold**, with $\Pi^t_{thres}(\boldsymbol{d})$ being the projected vector by restricting $\boldsymbol{d}$ to the dimensions j with values $\boldsymbol{d}_j \geq t$ (as in [12])
3. **Relative Threshold**, with $\Pi^t_{rel}(\boldsymbol{d})$ being the projected vector by restricting $\boldsymbol{d}$ to the dimensions j with values $\boldsymbol{d}_j \geq t * \boldsymbol{d}_{\alpha(1)}$, $t \in [0..1]$, thus restricting it to those values above a certain fraction of the highest-valued dimension
4. **Sliding Window**, with $\Pi^{t,l}_{window}(\Phi(d))$ being the projected vector by restricting $\boldsymbol{d}$ to the first i dimensions according to the order α_d for which the following condition holds: $\boldsymbol{d}_{\alpha(i-l)} - \boldsymbol{d}_{\alpha(i)} \geq t * \boldsymbol{d}_{\alpha(1)}$, $t \in [0..1]$ (as in the original ESA model [13])

A relevant question is certainly how to set the parameters m and t. We address this in the experiments by first fixing a reasonable value for m in Π^m_{abs}. In order to be able to compare the different approaches, we choose the parameter t in such a way that the number of non-zero dimensions of the projected ESA vectors of all documents in the datasets amounts to m on average. The parameter l was set to 100 as in [6].

3.3 Association Strength

In the following we will describe the different choices of the association strength function $as(d, a_i)$ between documents and articles determining the values of the ESA vector $\boldsymbol{d}$. These functions are based on the term vectors of d and a_i. As notation we use $|\mathcal{W}|$ as the number of articles, $|a_i|$ as number of terms in article a_i, $tf_d(w)$ ($tf_{a_i}(w)$) as the

[2] http://www.dmoz.org

term frequency of w in document d (article a_i), $rtf_{a_i}(w) = tf_{a_i}(w)/|a_i|$ as the relative term frequency and $af(w)$ as the number of articles containing term w in Wikipedia $\mathcal{W}$.

1. **TF.IDF**: The most widely used version of the $tf.idf$ function:
$$as_{tf.idf} := \sum_{w \in d} tf_d(w)\ rtf_{a_i}(w) \log \frac{|\mathcal{W}|}{af(w)}$$
2. **TF.IDF***: A modified $tf.idf$ version ignoring how often the terms occur in document d:
$$as_{tf.idf^*} = \sum_{w \in d} rtf_{a_i}(w) \log \frac{|\mathcal{W}|}{af(w)}$$
3. **TF** : An association function only based on term frequencies (ignoring inverse document frequencies):
$$as_{tf} = \sum_{w \in d} tf_d(w) rtf_{a_i}(w)$$
4. The **BM25** ranking function as defined by Robertson et al. [14] with parameters set to the following standard value: $k_1 = 2, b = 0.75$.
5. The **Cosine** similarity between the tf and $tf.idf$ vectors $\boldsymbol{d} = \langle tf_d(w_1), tf_d(w_2), \ldots \rangle$ and $\boldsymbol{a}_i = \langle tf.idf_{a_i}(w_1), tf.idf_{a_i}(w_2), \ldots \rangle$:
$$as_{cos} = \frac{< \boldsymbol{d}, \boldsymbol{a}_i >}{||\boldsymbol{d}|| \, ||\boldsymbol{a}_i||}$$

Note that we have also experimented with versions of the above where the tf_{a_i} instead of rtf_{a_i} values were used, yielding in all cases worse results with a performance degradation of about 75% in all cases. For this reason, we do not present the results with the tf_{a_i} versions of the above functions in detail.

3.4 Relevance Function

The relevance function $rel(q, d)$ defines the score of a document $d \in D$ for a given query q and is used to rank the documents in the retrieval process. In this multilingual setting, the function is defined on the translated and projected ESA vector $\hat{q} := \Pi(\Psi(\Phi(q)))$ of query q and the projected ESA vector $\hat{d} := \Pi(\Phi(d))$ of document d (see section 2.2).

Analogous to the Bag-of-Words model the ESA vectors can be seen as Bag-of-Articles model for a document d. The term frequency of $a_i \in \mathcal{W}$ is defined as $tf_d(a_i) := \hat{d}_{a_i}$, the document frequency $df(a_i)$ is the number of documents in D with $\hat{d}_{a_i} > 0$. Based on this model different relevance functions defined for text retrieval can by applied to the ESA vectors.

- The **Cosine** similarity of query and document vectors (used by all ESA implementations known to us):
$$rel_{Cosine} = \frac{< \hat{q}, \hat{d} >}{||\hat{q}|| \, ||\hat{d}||}$$

- **TF.IDF:** The TF.IDF function transfered to the Bag-of-Articles model:

$$\begin{aligned} rel_{tf.idf} &= \sum_{a \in \mathcal{W}} tf_q(a) rtf_{d_i}(a) idf(a) \\ &= \sum_{a \in \mathcal{W}} \hat{q}_a \frac{\hat{d}_a}{\sum_{a^* \in \mathcal{W}} \hat{d}_{a^*}} \log \frac{|D|}{df(a)} \end{aligned}$$

- **KL-Divergence:** Many recent text retrieval systems use relevance functions based on the theory of language modeling. In order to be able to apply these approaches to our setting we define the conditional probability of an article given a document as follows:

$$P(a|d) := \frac{\hat{d}_a}{\sum_{a^* \in \mathcal{W}} \hat{d}_{a^*}}$$

This definition of the conditional probability originates from the bag-of-words model and is inspired by [15], where it is also described how these probabilities can be used to define a ranking function based on the Kullback-Leibler divergence [16], which measures the difference between the query and the document model (leading ultimatively to the negative sign in the formula below). Transferred to our model this results in the following retrieval function:

$$rel_{KL} = -D_{KL}(q \| d) \cong -\sum_{a \in \mathcal{W}} P(a|q) \log P(a|d)$$

- **LM:** An alternative approach is to use the conditional probability $P(q|d)$ as relevance function. This distribution can be converted using the conditional distributions of documents given articles, Bayes law and the a priori probability of articles $P(a) = \frac{df(a)}{|D|}$:

$$\begin{aligned} rel_{LM} = P(q|d) &= \sum_{a \in \mathcal{W}} P(q|a) P(a|d) \\ &\cong \sum_{a \in \mathcal{W}} \frac{P(a|q)}{P(a)} P(a|d) \end{aligned}$$

4 Experiments

Our experiments have been carried out in an iterative and greedy fashion in the sense that we start form the original ESA model as a baseline, then iteratively varying different parameters and always fixing the best configuration before studying the next parameter. At the end of our experiments we will thus be able to assess the combined impact of the best choices on the performance of the ESA model.

To prove the significance of the improvement of our best settings (projection function Π^{10000}_{abs}, association strength function TF.IDF*, cosine retrieval model) we carry out paired t-tests (confidence level 0.01) comparing the best settings pairwise with all other results for all language pairs on both datasets. Results where the differences are **not** significant with respect to **all** other variants at a confidence level of 0.01 are marked with "X" in Figure 1 to 4.

4.1 Datasets and Evaluation Measures

For the ESA implementation we used the English, German and French Wikipedia database[3]. As we rely on the language links to map the ESA vectors to other languages, we only chose articles that are linked across all three languages. This means that the mapping function $m_{a \to b}$ used for CL-ESA is defined for all articles and is a bijection between the Wikipedia subsets for all language pairs considered. Altogether we used 166,484 articles in every language.

To evaluate the performance of the CLIR system we performed mate retrieval on two well known parallel corpora: The Multext corpus derived from the Multext project[4] consisting of 2,783 question and answer pairs, and the JRC-ACQUIS corpus[5] consisting of 15,464 documents. For both datasets documents in one language were taken as queries to search the documents in another language. In this case automatic evaluation is possible as the relevant document, i.e. the translation of the query, is known in advance. All mentioned collections were prepared using common IR-like preprocessing steps including elimination of stopwords, special characters and extremely short terms (length < 3) and stemming.

As evaluation measures we used TOP-k accuracy, i.e. the number of queries for which the mate was found in the top k documents, and Mean Reciprocal Rank, which measures the average position of the mate documents (all standard measures in information retrieval). As the observed effects were constant across measures, we only present TOP-1 accuracy in Figures 1 to 4. For experiments on the Multext corpus we used all documents (2,783) as queries to search in all documents in the another languages. The results for language pairs were averaged for both retrieval directions (e.g. using English documents as queries to search in the German documents and vice versa). For the JRC-ACQUIS dataset we randomly chose 3000 parallel documents as queries (to yield similar settings as in the MULTEXT scenario) and the results were again averaged for language pairs. This task is harder compared to the experiments on the Multext corpus as the search space now containing 15,464 documents is bigger by a factor of approximately 5, which explains the generally lower results on the JRC-Acquis dataset.

4.2 Results

In the following we discuss the results of the different variations of the CL-ESA model:

Projection Function. We first used different values for the parameter m in the projection function Π^m_{abs}. The results in Figure 1 showed that $m = 10,000$ is a good choice for both datasets.

On the basis of this result, we investigated different projection functions. In order to be able to compare them, we set the different threshold values t such that the projected

[3] Snapshot of 03/12/2008 (English), 06/29/2008 (German) and 06/25/2008 (French).

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

[5] http://wt.jrc.it/lt/Acquis/

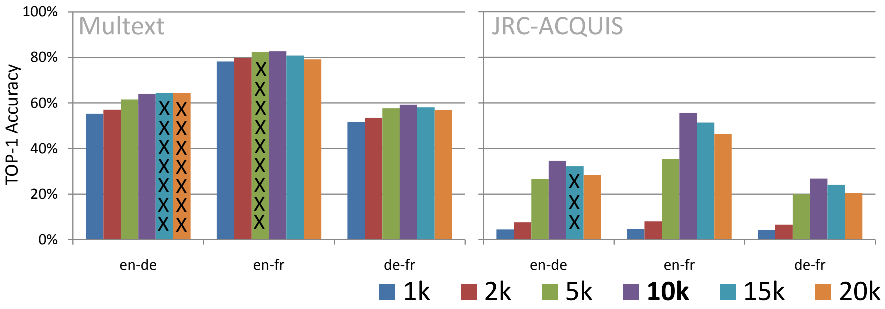

Fig. 1. Variation of m in Π_{abs}^{m} using the TF.IDF* association function and cosine retrieval model

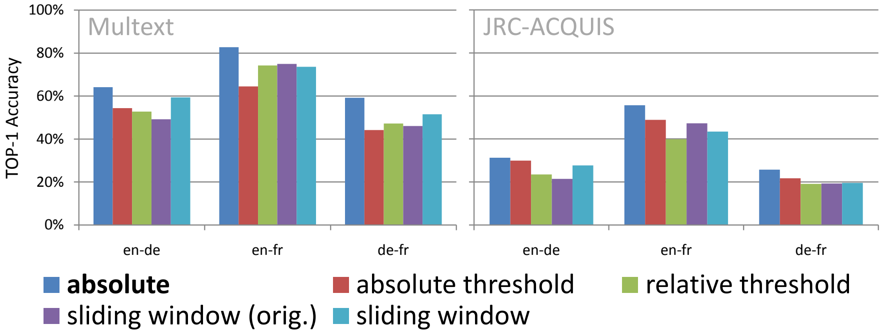

Fig. 2. Variation of the projection function Π using the TF.IDF* association function and cosine retrieval model

ESA vectors had an average number of approx. 10,000 non-zero dimensions. An exception is the function *sliding window (orig.)* where we used the parameters described in [13]: $t = 0.05$ and $l = 100$. Using an absolute number of non-zero dimensions yielded the best results (see Figure 2), the difference being indeed significant with respect to all other variants. Thus, we conclude that neither the settings of the original ESA approach (sliding window) nor in the model of Gurevych et al. (fixed threshold) are ideal in our experimental settings. For the remaining experiments we thus fix the absolute dimension projection function with 10,000 articles ($\Pi_{abs}^{10,000}$).

Association Strength. The results in Figure 3 show that the functions TF.IDF (used in the original ESA model) and TF.IDF* perform much better compared to the other functions. The better performance of TF.IDF*, which ignores the term frequencies in the queries, was indeed significant w.r.t. all other alternatives for all language pairs considered on both datasets. We thus conclude that the settings in the original ESA model are reasonable, but, surprisingly, can be improved by ignoring the term frequency of the words in the document to be indexed. The low results using the TF function show that IDF is an important factor in the association strength function. Otherwise the normalization of the TF.IDF values (= Cosine function) reduces the retrieval performance substantially.

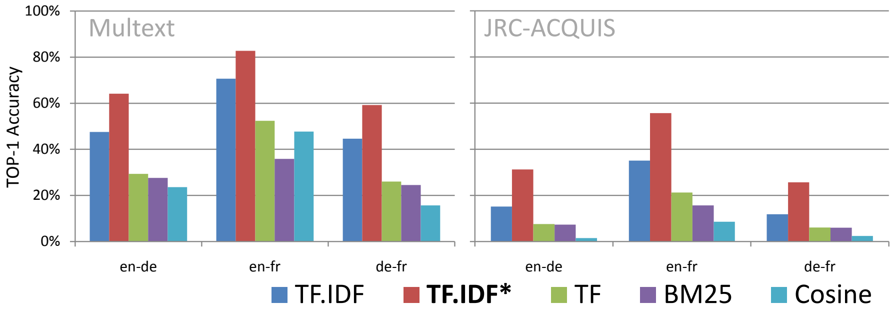

Fig. 3. Variation of the association strength function as using the projection function $\Pi_{abs}^{10,000}$ and cosine retrieval model

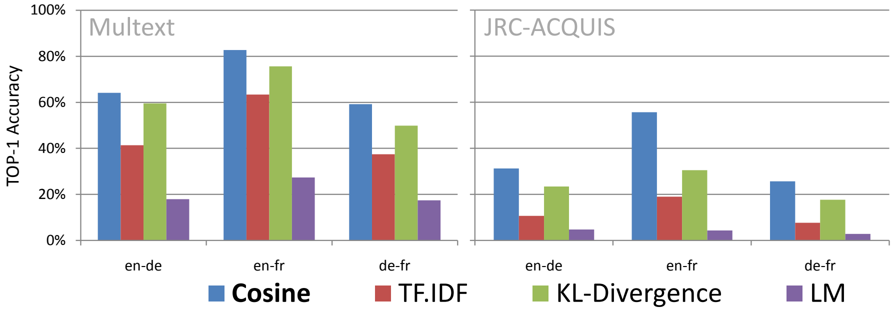

Fig. 4. Variation of the retrieval model using $\Pi_{abs}^{10,000}$ and TF.IDF*

Retrieval Model. The variations of the retrieval model lead to the result that the cosine function, which is used by all ESA implementations known to us, constitutes indeed a reasonable choice. All other models perform worse (the difference being again significant for all language pairs on both datasets), which can be seen at the charts in Figure 4, especially on the JRC-ACQUIS dataset.

4.3 Discussion

Our results show on the one hand that ESA is indeed quite sensitive to certain parameters (in particular the association strength function and the retrieval model), the choices for which can have a large impact on the performance of the approach. For example, using a tf_{a_i} values instead of rtf_{a_i} (which is length normalized) values in the association strength function decreases performance by about 75%. Unexpectedly, abstracting from the number of times that a word appear in the query document (using TF.IDF*) improves upon the standard TF.IDF measure (which takes them into account) by 17% to 117%. We have in particular shown that all the settings that are ideal in our experiments are so indeed in a statistically significant way (with the exception of the number of dimensions taken into account).

On the other hand, while we can confirm by our experiments that the settings in the original ESA model ($\Pi_{window}^{0.05,100}$, TF.IDF, cosine) [6,13] are reasonable, it is also the case that with the settings which according to our experiments are ideal on both datasets ($\Pi_{abs}^{10,000}$,$TF.IDF^*$,cosine) we achieve a relative improvement in TOP-1 accuracy between 62% (from 51.1% to 82.7%, Multext dataset, English/French) and 237% (from 9.3% to 31.3%, JRC-ACQUIS dataset, English/German), which shows again that the settings can have a substantial effect on the ESA model and that ESA shows the potential to be further optimized and yield even better results on the various tasks it has been applied to.

Finally, all experiments including the German datasets have worse results compared to the English/French experiments. This is likely due to the frequency of specific German compounds in the datasets, which lead to a vocabulary mismatch between documents and Wikipedia articles. However an examination of this remains for future work.

5 Research Context and Conclusion

We have mentioned already different approaches for folding in "semantics" (meaning very different things depending on the approach in question) into information retrieval tasks (see Section 1). We have examined in particular the ESA model in this paper, which has gained substantial attention in recent years [17,3,9,8] since it was originally published in 2007 [6] and partially already (not under this name) in 2005 [18]. The original application of the ESA model was the computation of semantic relatedness between words. In fact, Gabrilovich and Markovitch showed that the ESA model outperforms bag-of-word and latent semantic indexing approaches on this task. ESA has been also exploited in text classification approaches [18,17,19] where it has been already shown that an appropriate dimension selection function has significant influence on the performance of the ESA model. ESA has been also applied with reasonable success to information retrieval settings [3], in particular cross-language retrieval settings [9,8]. In this paper we have generalized the original ESA model and made explicit the degrees of freedom that it offers and highlighted the different choices that various implementations have adopted. The starting point for our investigation has been the observation that none of the above works has examined the various possible choices systematically due to the fact that they have focused on different aspects and this was not their main research question. In any case, if the ESA model continues to be applied successfully to various text-centered tasks, a systematic investigation of the impact of different choices seems definitely necessary. We have provided such an analysis in the context of a cross-lingual mate retrieval task (presenting results on two datasets), showing which choices have or don't have a large impact and confirmed that the settings of the original ESA model are indeed reasonable, something which to our knowledge has never been shown, but can still be improved for cross-lingual retrieval settings. Our results are clearly limited to the type of cross-lingual mate retrieval task that we have considered and an avenue for future work could be the investigation of the choices under consideration for monolingual or more general (non-mate-retrieval like) cross-lingual retrieval tasks, text classification or semantic relatedness computation. The examination of optimal settings constitutes an interesting topic for future investigation which can help to shed additional light on the ESA approach. Furthermore, our generalization of ESA can help to guide such investigations in the future, providing a common framework for comparisons.

Acknowledgments

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

References

1. Richardson, R., Smeaton, A.: Using wordnet in a knowledge-based approach to information retrieval. In: Proceedings of the BCS-IRSG-Colloquium (1995)
2. Schütze, H., Pedersen, J.: A cooccurrence-based thesaurus and two applications to information retrieval. Information Processing and Management 33(3), 307–318 (1997)
3. Gurevych, I., Müller, C., Zesch, T.: What to be? - electronic career guidance based on semantic relatedness. In: Proceedings of ACL (2007)
4. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society of Information Science 41(6), 391–407 (1990)
5. Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J.: Indexing with wordnet synsets can improve text retrieval. In: Proceedings of the COLING/ACL 1998 Workshop on Usage of WordNet for NLP, pp. 38–44 (1998)
6. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of IJCAI, pp. 1606–1611 (2007)
7. Furnas, G., Landauer, T., Gomez, L., Dumais, S.: The vocabulary problem in human-system communication. Communications of the ACM 30(1), 964–971 (1987)
8. Sorg, P., Cimiano, P.: Cross-lingual information rerieval with explicit semantic analysis. In: Working Notes of the Annual CLEF Meeting (2008)
9. Potthast, M., Stein, B., Anderka, M.: A wikipedia-based multilingual retrieval model. In: Proceedings of ECIR, pp. 522–530 (2008)
10. Littman, M., Dumais, S., Landauer, T.: Automatic Cross-Language Information Retrieval using Latext Semantic Indexing. In: Cross-Language Information Retrieval, pp. 51–62. Kluwer, Dordrecht (1998)
11. Dumais, S., Letsche, T., Littman, M., Landauer, T.: Automatic cross-language retrieval using latent semantic indexing. In: Proceedings of the AAAI Symposium on Cross Language Text and Speech Retrieval (1997)
12. Müller, C., Gurevych, I.: Using wikipedia and wiktionary in domain-specific information retrieval. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V. (eds.) Evaluating Systems for Multilingual and Multimodal Information Access. LNCS, vol. 5706, pp. 219–226. Springer, Heidelberg (2009)
13. Gabrilovich, E.: Feature Generation for Textual Information Retrieval using World Knowledge. PhD thesis, Israel Institute of Technology, Haifa (2006)
14. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at trec-3. In: Proceedings of TREC (1994)
15. Zhai, C.X., Lafferty, J.D.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of CIKM, pp. 403–410 (2001)

16. Lee, L.: Measures of distributional similarity. In: Proceedings of ACL (1999)
17. Egozi, O., Gabrilovich, E., Markovitch, S.: Concept-based feature generation and selection for information retrieval. In: Proceedings of AAAI (2008)
18. Gabrilovich, E., Markovitch, S.: Feature generation for text categorization using world knowledge. In: Proceedings of IJCAI (2005)
19. Gupta, R., Ratinov, L.: Text categorization with knowledge transfer from heterogeneous data sources. In: Proceedings of AAAI, pp. 842–847 (2008)