Combining Wikipedia-Based Concept Models for Cross-Language Retrieval

Benjamin Roth and Dietrich Klakow

Spoken Language Systems, Saarland University, Germany
beroth@coli.uni-saarland.de,dietrich.klakow@lsv.uni-saarland.de

Abstract. As a low-cost ressource that is up-to-date, Wikipedia recently gains attention as a means to provide cross-language brigding for information retrieval. Contradictory to a previous study, we show that standard Latent Dirichlet Allocation (LDA) can extract cross-language information that is valuable for IR by simply normalizing the training data. Furthermore, we show that LDA and Explicit Semantic Analysis (ESA) complement each other, yielding significant improvements when combined. Such a combination can significantly contribute to retrieval based on machine translation, especially when query translations contain errors. The experiments were perfomed on the Multext JOC corpus und a CLEF dataset.

Key words: Latent dirichlet allocation, explicit semantic analysis, cross-language information retrieval, machine translation

1 Introduction

Dimensionality reduction techniques have traditionally been of interest for information retrieval as a means of mitigating the word mismatch problem. The term *concept model* is more general than dimensionality reduction and denotes a mapping from the word space to another representation that provides a smoother similarity measure than word statistics and is often induced from co-occurrence counts on paragraph or document level. Such a representation may for example be obtained by matrix approximation [7], by probabilistic inference [20] or techniques making use of the conceptual structure of corpora such as Wikipedia [9].

Cross-language information retrieval can be viewed as an extreme case of word mismatch, since for any two texts the vocabulary is in general disjoint if the languages are not the same. In order to have a cross-lingual similarity measure, it is necessary that concept spaces of different languages are aligned, which is often achieved by extending the notion of co-occurrence to pairs of translated or thematically related texts. While some work has been done on multilingual concept modeling [5, 8, 15–19], often the focus is on one method and a comparison with other methods is missing (but see [4] for a comparison of classical cross-language retrieval models). One reason for this might be that concept models require adaptations for multilinguality that do not seem to be easily implemented. We will show that the adaptations can in fact be minimal and on the data side only. Another question that has not been investigated so far is how different multilingual concept models can contribute to each other and how they can be combined with word models, an approach that is standard for the monolingual case.

The rest of the paper is structured as follows: In section 2 we summarize standard and multilingual Latent Dirichlet Allocation, a probabilistic concept model, and Explicit Semantic Analysis, a more recent explicit concept model. In section 3 we outline how we think the forementioned methods can be applied more profitably. Our experiments are described in section 4. In section 4.1 we explore both LDA and ESA on a mate retrieval task and observe much better results for LDA than reported so far and obtain consistent improvement for their combination. In section 4.2 we show how concept models can improve word-based cross-language retrieval. We end with an outlook on future work and conclusion.

2 Related Work

2.1 Latent Dirichlet Allocation

Probabilistic Model Latent Dirichlet Allocation (LDA) [2, 10, 20] is a latent variable model that gives a fully generative account for documents in a training corpus and for unseen documents. Each document is characterized by a topic distribution, words are emitted according to an emission probability dependent on a topic. The main difference to pLSA [11, 12] is that both topic distributions and word emission distributions are assumed to be generated by Dirichlet priors. It is common to parameterize the Dirichlet prior uniformly with parameters $\alpha = \alpha_1 = \cdots = \alpha_T$ and to direct only the “peakiness” of the multinomials drawn from it. The LDA model describes the process of generating text in the following way:

1. For all k topics generate multinomial distributions $\psi^{(z_k)} = p(w_j|z_k) \sim Dir(\beta)$.
2. For every document d:
   (a) Generate a multinomial distribution $\theta^{(d)} = p(z_k|d) \sim Dir(\alpha)$.
   (b) Generate a document length, and topics $z_i \sim \theta^{(d)}$ for every position i in the document.
   (c) Generate words $w_i \sim \psi^{(z_i)}$ for every position in the document.

Usually, no generative account for the length of the document is given.

Practical Issues The first approach to estimate such a model [2] was to represent and estimate ψ and θ explicitly, resulting in different inference tasks to be solved and combined. Later approaches concentrate on getting a sample of the assignment of words to topics instead by Gibbs sampling [10].

To determine the similarity between two documents, one can compare either their sampled topic vectors or the probability vectors obtained from them

[10]. When other variational EM estimation techniques are applied, also other parameter vectors might be available and used [2, 5]. The comparison between these vectors can be done either by taking the cosine similarity of their angles or by using probability divergence measures. For language model based information retrieval, one is interested in the probability of a query, given a document. Wei and Croft [22] interpolate a language model based on LDA with a unigram language model directly estimated on the document.

Whenever a sampling technique is used, one wants to be sure that the estimates are stable. This could be a problem when only one topic is sampled per position for short documents or queries. The most natural way to overcome this problem is to average the results of several sampling iterations.

Multilingual LDA LDA has been extended for several languages [16] (see also [15] for an investigation of the semantic clustering of this model and [6] for cross-lingual news-linking with it). Most of the components remain the same, the main difference is that for each language l_s a different word emission distribution ψ_{l_s} is assumed. Depending on the language of a position in a document, a word is generated conditioned on the corresponding distribution. The model does not use the topic variable to estimate the language, as it could be the case for a monolingual model applied to a multilingual document without any adaptions on either the data or the model side. A theoretically sound model does not mean that it also provides a good bridging between two languages. It is crucial [15] how many “glue documents”, i.e. documents that indeed have counterparts in all compared languages, are available: Although the model does not try to capture the language, the latent variables might tend to structure the monolingual spaces without semantic alignment between the languages (imagine a multilingual text collection with only one glue document as an extreme case).

2.2 Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) [5, 9, 17–19] is another scheme to overcome the word-mismatch problem. In ESA, the association strength of words to the documents in a training collection is computed, and vectors of these asscociations are used to represent the words in the semantic space spanned by the document names. These word representations can be used to compare words, they can also be combined for a comparison of texts. See [1] for the relation to the Generalized Vector Space Model [23].

Formalization Several formalizations are possible in this setting. The fundamental ingredients that determine an implementation are:

- **Word vectors:** For every word w a vector $\mathbf{w}$, indicating its association strength to the documents is computed.
- **Text vectors:** For a new document (or query) d a vector representation $\mathbf{d}$ is computed from the word vectors.

- **Similarity function:** Giving two documents d_1 and d_2 the similarity is computed using a similarity function on their text vectors.

The word vectors that were used in [9] and found to be optimal in [19] are obtained by taking the respective columns of the tf.idf-weighted document-term matrix A of the training collection. We use this choice of word vectors in our experiments.

For the text similarity, several settings have been proposed. In [9] a weighting of the word vectors is used, they are multiplied with scalars equal to, again, an tf.idf weighting of the terms, and then summed. Sorg and Cimiano [19] explore further combination schemes, including the sum of the elements of either the multiset (considering term frequency) or of the set (not considering term frequency) of word vectors. They find that the set combination works best, yielding preliminary text vectors of the form:

$$\hat{\mathbf{d}} = \sum_{w \in d} \mathbf{w}$$

It is beneficial to truncate the resulting vectors at a certain threshold. The thresholding that turned out to be most successful [19] was to retain the 10000 biggest non-zero values of the vectors $\hat{\mathbf{d}}$. Again, we use this parametrization in all following experiments that involve ESA. As a similarity function the cosine is suggested [19] and used by us.

Multilingual ESA The application of this model in a multilingual setting is straightforward. For L languages consider document term matrices $A^{(1)} \cdots A^{(L)}$. Construct the matrices in a way that the document rows correspond. For all languages each of the rows $A_n^{(\cdot)}$ contains documents about the same topic across the languages. Therefore only documents can be included that are available in all of the considered languages. For each document the mapping to text vectors is performed using a monolingual matrix corresponding to its language. As the documents are aligned, similarities can be computed across languages. Because the relative frequency is used in the tf.idf-weighting, all documents are normalized and no bias occurs for documents longer in one language than in another.

3 Making Use of Concept Models for CLIR

Making Use of LDA In our experiments we want to generalize the intuition given for the multilingual LDA model [15]: not only should a large number of glue documents exist, good bridging documents should optimally be of equal length. Experimentation with a small fraction of the training data indicated that the multilingual LDA model and a monolingual LDA model on documents normalized to equal length on both language sides yield about the same performance, while a monolingual model on unnormalized data performs considerably worse. Moreover, highly optimized and parallelized toolkits that allow us to perform training on all Wikipedia articles have only been developed for standard LDA.

We believe, therefore, that it is a promising approach to normalize the data suitably to be processed with a standard monolingual LDA model.

Wikipedia is used as a parallel training corpus: corresponding articles are concatenated, their length is normalized to match the length of the counterpart in the other language. We propose two methods of length normalization: First, to cut off every document at a certain length. Second, to retain for the longer language side of an article only a random sample of size equal to the smaller language side. A resizing with a scalar is not possible because the sampling process requires integer counts.

The vocabulary is uniquely identified for every language by attaching suitable prefixes (*en_*, *de_*) to the words. Similarity is measured between two texts after inference by taking the cosine between their vectors of sampled topic statistics.

Making Use of ESA ESA is applied as described in section 2.2, which is as closely as possible as reported in [19].

Making Use of Machine Translation The machine translation retrieval model translates the queries with a standard Moses [14] translation model trained on Europarl [13]. Translated queries and text are then compared by the cosine of their tf.idf-weighted word-vectors.

For the document-term matrix D of the target collection with N documents, we use the commonly used weighting function

$$tf.idf(w,n) = \frac{D_{n,w}}{\sum_{w' \in W} D_{n,w'}} \log \frac{N}{\sum_{n'=1}^{N} \mathbf{1}_{D_{n',w}>0}}$$

Here, $\mathbf{1}_{D_{n',w}>0}$ is an indicator that equals to one if word w has appeared in document $d_{n'}$ at least once, and that equals to zero otherwise.

Model Combination We use a simple scheme to combine models by concatenating L_2-normalized vectors. Let $\mathbf{u}$ be a m-dimensional vector and $\mathbf{v}$ be a n-dimensional vector which represent the same document in two different models. Then the model combination with interpolation weight α represents this document by an $m+n$-dimensional vector $\mathbf{w}$:

$$\mathbf{w_i} = \begin{cases} \alpha \frac{\mathbf{u_i}}{|\mathbf{u}|} & \text{if } 1 \leq i \leq m \ , \\ (1-\alpha)\frac{\mathbf{v_{(i-m)}}}{|\mathbf{v}|} & \text{otherwise.} \end{cases}$$

This way any models can be combined as long as they are in vector representation. This the case for all models mentioned above, although they rely on very different principles. Similar combinations have been proven effective in the case of pLSA for monolingual retrieval [11].

4 Experiments

We are using two datasets, the Multext JOC corpus for the task of finding translations, and a CLEF 2000 query-based retrieval collection. These datasets and the experiments are described in detail in the next section.

4.1 Mate Retrieval on Multext JOC

There is only one publication [5] known to us that compares LDA for cross-language information retrieval with ESA. Interestingly, our experiments on the same dataset will suggest a distinctively different assessment of the potential of LDA for the same task.

The basis of the evaluation is the Multext JOC corpus[1] which consists of 3500 questions to the European Parliament and of answers to these questions. As in [5] we use the concatenation of a question together with its answer as a query in one language to search the collection of translations in another language for its counterpart. Our experiments were done with English as the query language and German as the target language. Only preprocessing steps that are clear and easy to reproduce were performed. Exactly those questions were retained that to which an answer was assigned and had the same id in English and German. This resulted in a set of 3212 texts in each language, 157 more than were used in [5][2]. Sequences of characters in Unicode letter blocks were considered words. Words with length $= 1$ or length > 64 and words contained in the Snowball stopword list were ignored. All other words were stemmed with the publicly available Snowball stemmer[3]. In contrast to [5], no compound splitting was done.

For the training collection all pairs of Wikipedia articles[4] were used that have bidirectional direct cross-language references. All markup was stripped off by using the same filter as in a publicly available ESA implementation[5]. Wikipedia articles of less than 100 words in either language were ignored and words with a Wikipedia document frequency of 1 were filtered out. The final training corpus consists of 320000 bilingual articles.

Performance of retrieval was measured in mean reciprocal rank (mrr). The ESA retrieval experiment was performed using the same parametrization as discribed before and the result of [5] was reproduced to a difference of 1% (in our experiments we obtained a score of $mrr = 0.77$ compared with $mrr = 0.78$).

As for the LDA experiments, we were interested in the effect of length normalization of the training documents. We compare two methods: First, every document was cut off at a length of 100 words. Second, the method of down sampling the longer language side to the length of the smaller one was applied. We marked each word with a prefix indicating its language and retained a vocabulary size of roughly 770 thousand and 2.1 million for the cut-off method and for

[1] http://www.lpl.univ-aix.fr/projects/multext

[2] the exact document selection criterion of their experiments is unknown to us

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

[4] we used the German snapshot of 2009-07-10 and the English snapshot of 2009-07-13

[5] http://code.google.com/p/research-esa/

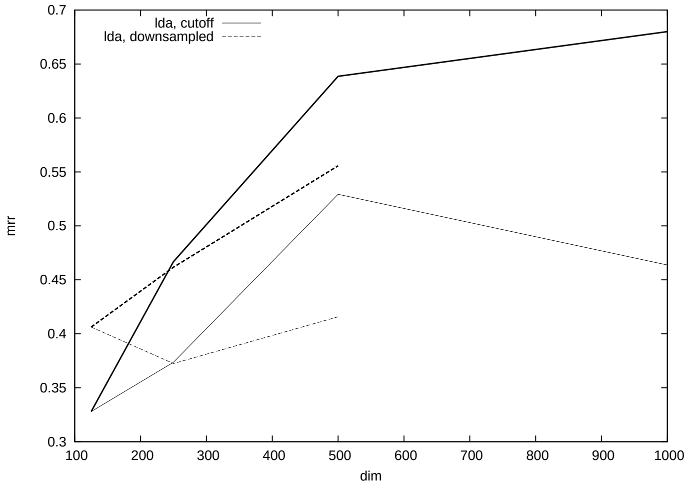

Fig. 1. Performance of LDA models estimated with dimensions (=numbers of topics) equal to 125, 250, 500 and 1000. Thick lines indicate combinations of all models up to the dimension on the $x-$axis. Drops in the perfomance curve on single points may be due to local sampling optima.

the downsampling method respectively. Both training collections were embedded with 125, 250 and 500 dimensions, and additionally with 1000 dimensions for the cut-off corpus (the vocabulary size was the limiting factor with respect to our computing facilities). The Google plda package [21] was used with the suggested parameters ($\alpha = \frac{50}{\#\text{topics}}$ and $\beta = 0.01$). With the trained model, topics were inferred for the monolingual Multext documents. In order to get a stable estimate, the statistics of 50 sampling iterations were averaged. Similarity in the LDA setup was measured by taking the cosine similarity between the sampling statistics.

Table 1. Performance of LDA on Multext

LDA method	**number of topics**	**mrr**
Cimiano et al.	500	.16
length downsampling	500	.42
length cut-off	500	$.53^{**}$
length downsampling	$125 + 250 + 500$	$.55^{**}$
length cut-off	$125 + 250 + 500 + 1000$	$.68^{**}$

A drastic improvement over non-normalized LDA can be observed: while [5] report a score of $mrr = 0.16$ for their 500-dimensional LDA model, we get $mrr = 0.53$ with the cut-off corpus. We suppose that the reason for this difference is that a non-multilingual LDA model applied to a comparable corpus

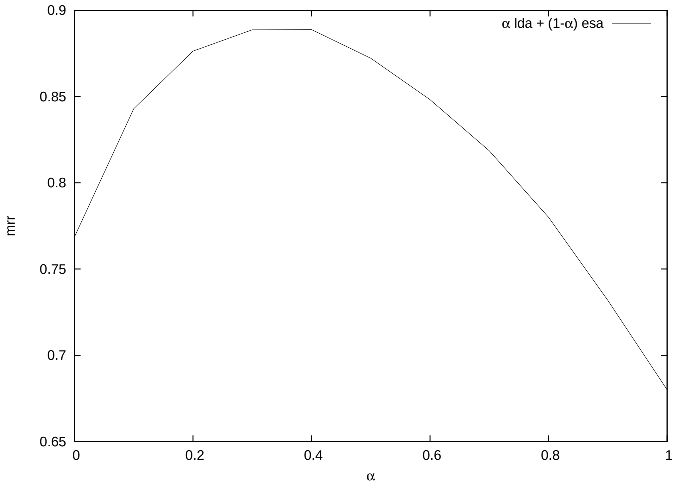

Fig. 2. Combining LDA and ESA on the Multext corpus. The improvement over ESA alone is significant with $p \ll 0.005$ for $.1 \leq \alpha \leq .7$.

estimates the predominant language of a document rather than its semantic content. Another improvement can be observed by combining the results of different models, a technique that is usually applied for pLSA [12]. In this case, the samping statistics of runs with different dimensional models were L_2-norm normalized and concatenated without further weighting. This yielded a score of $mrr = 0.68$ for the cut-off model, showing performance in the same order of magnitude as ESA. Figure 1 and Table 1 give a survey of the results obtained with LDA. Scores significantly better than in the respective line above having $p \ll 0.005$ in the paired t-test are marked with $**$. (Of course we could not test against scores reported in [5], for lack of the original numerical data.)

In order to determine how different the ESA and the LDA models are and how much can they contribute to each other, we combined the vector representations of both models by different interpolation factors $0 \leq \alpha \leq 1$. A stable improvement in performance with maximum $mrr = 0.89$ was achieved for giving the cut-off LDA model a weight of 0.4 and the ESA model a weight of 0.6. See Figure 2.

4.2 Query-Based Retrieval with CLEF2000

Mate retrieval experiments can be criticized as being an unrealistic retrieval scenario. Therefore, a second evaluation was done on the CLEF[6] German-English ad-hoc track of the year 2000. The target corpus consists of about 110000 English

[6] See `http://www.clef-campaign.org/`; The evaluation packages are available via the ELRA catalogue `http://catalog.elra.info`, The CLEF Test Suite for the CLEF 2000-2003 Campaigns, catalogue reference: ELRA-E0008.

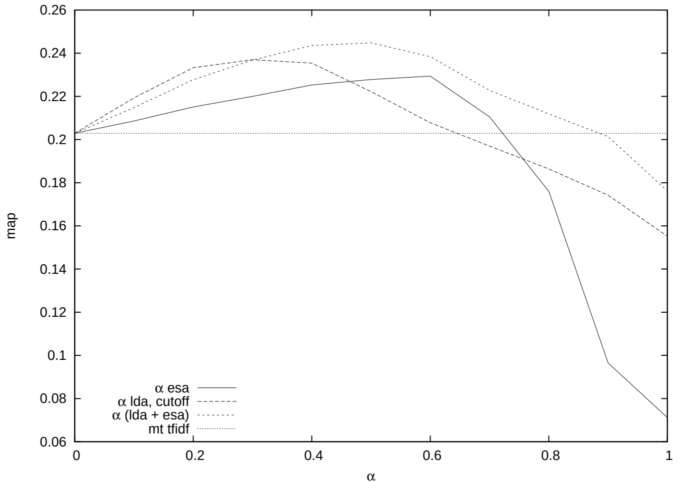

Fig. 3. Interpolation of concept models (having interpolation weight α) with machine translation (weighted $1 - \alpha$) on CLEF2000. The improvement given by the LDA-ESA is significant with $p < 0.005$ for $.2 \leq \alpha \leq .6$.

newspaper articles together with 33 German queries for which relevant articles could be pooled. For our experiments the title and description fields of the queries were used and the narrative was ignored.

A common strategy for cross-language retrieval is first to translate the query and then to perform monolingual retrieval. While the translation process would have taken prohibitively long for the Multext corpus, we performed query translation on the CLEF2000 queries with a standard Moses translation model trained on Europarl. Retrieval with the translated queries was done by comparing the cosine of the tf.idf-weighted word-vectors.

We evaluated both the machine translation model and the concept models trained on Wikipedia. In addition to the most commonly used mean average precision score (map) we also evaluated by geometric mean average precision (gmap), which rewards stable results for hard queries. The ESA and the cut-off LDA models with $dim = 500$ perform equally well for *map*, while the combination of LDA dimensions gets a considerably better score. This is in contrast to the findings in the mate-retrieval setup. The reason for that, we suspect, may be that the parameters of ESA have been found in order to optimize such a setting.

For *gmap*, LDA consistently outperforms ESA. For the combined LDA-ESA model a slight improvement could be observed using the combination with $\alpha = .5$ (henceforth referred to as *combined concept model*). The machine translation model ($map = .203$) performed better than LDA and ESA. When interpolated with the machine translation model all three concept models (LDA, ESA and combined) achieved improvements. The biggest and most stable improvement was achieved by the interpolation of machine translation and combined concept model yielding a score up to $map = .245$ with equal weight for the concept model and the machine translation model. Figure 3 and Table 2 show an overview of the results. Scores significantly better than in the respective line above, having $p < 0.05$ and $p < 0.005$ in the paired t-test, are marked with $*$ and $**$ respectively. The evaluation results of the combined system lie well within the range of the participating systems in CLEF 2000 for the same track[7]. Particularly, no manually edited lexicon and no compound-splitter is used in our case.

Table 2. Query-based retrieval on CLEF2000

method	parameters	map	gmap
ESA		.071	.003
LDA, cut-off	$d = 500$	.071	.010
LDA, cut-off	$d = all$	$.155^*$	$.043^*$
LDA+ESA	$\alpha = .5$	.176	.054
MT tf.idf		.203	.061
concept+MT	$\alpha = .5$	$.245^{**}$	$.128^{**}$

Error Analysis A querywise error analysis is difficult as the inner workings of quantitative methods are often opaque to human analysis. However, the machine translation output is the most contributing source and it is accessible to examination. We sorted the machine translation output by how much it profited by the concept models in the best performing setting. In Table 3 we report the score that is obtained by machine translation and the increase when combined with the concept models. We analyzed how often a word was obviously unknown by the machine translation system trained on Europarl and therefore wrongly just copied over. It would be possible to recognize this type of error automatically. In addition, for every translated query we counted how many words in it had no semantic meaning related to the purpose of the query and were therefore useless (these words are hence called *junk words*). Junk words are, for example, function words not filtered by the stopword list, machine translation errors of several kinds and artefacts from the query formulation (e.g. *"Gesucht sind Dokumente, die ... liefern"* in the description part of query 4). The junk word error type would be more difficult to detect.

Although the analyzed data basis is small, we conjecture that the concept model makes such queries more robust which induce one of the two errors, while it might be less useful where a good translation is present and the terms are weighted well: In the cases where the concept models contributed there were, on average, 0.53 unknown words for machine translation and 0.24% junk words, in contrast to 0.14 unknown words and 0.04% junk words in the cases where

[7] For licensing reasons, we are not allowed to make a direct comparison, but see [3] for a survey.

Table 3. Improvement of *map* through the ESA+LDA concept component compared with error rates. The 4th column indicates the change in comparison to machine translation alone (3rd column). The 5th column contains one + per unknown word, the last column the percentage of junkwords per query.

		scores		errors	
id	**query title**	**mt *ap***	Δap	**unknowns**	**junk(%)**
18	Unfälle von Brandbekämpfern	0.000	$ap = 0.033$	+	0.71
9	Methanlagerstätten	0.000	$ap = 0.017$	++	0.50
7	Doping und Fußball	0.004	+361.36%		0.33
4	Überschwemmungen in Europa	0.007	+209.09%		0.17
26	Nutzung von Windenergie	0.071	+167.40%		0.13
11	Neue Verfassung für Südafrika	0.093	+111.95%		0.09
12	Sonnentempel	0.009	+109.57%	+	0.50
17	Buschbrände bei Sydney	0.138	+96.09%	++	0.29
5	Mitgliedschaft in der Europäischen Union	0.060	+89.38%		0.08
15	Wettbewerbsfähigkeit der europäischen...	0.177	+88.51%		0.10
38	Rückführung von Kriegstoten	0.045	+86.84%	+++	0.33
21	Europäischer Wirtschaftsraum	0.165	+73.38%		0.22
14	USA-Tourismus	0.015	+67.72%	+	0.17
28	Lehrmethoden für nicht-englischsprachige...	0.072	+61.43%	+	0.29
33	Krebsgenetik	0.382	+52.46%	+	0.38
16	Die Französische Akademie	0.084	+46.84%		0.33
13	Konferenz über Geburtenkontrolle	0.137	+46.68%		0.17
40	Privatisierung der Deutschen Bundesbahn	0.015	+46.10%		0.25
31	Verbraucherschutz in der EU	0.222	+46.03%		0
32	Weibliche Priester	0.234	+34.99%		0.22
20	Einheitliche europäische Währung	0.218	+24.13%		0
36	Produktion von Olivenöl im Mittelmeerraum	0.269	+23.29%		0.11
22	Flugzeugunfälle auf Start- und Landebahnen	0.094	+19.76%	+	0.22
37	Untergang der Fähre Estonia	0.926	+4.13%		0.17
19	Golfkriegssyndrom	0.704	+1.44%		0.25
1	Architektur in Berlin	0.651	+1.27%		0.17
30	Einsturz einer Supermarktdecke in Nizza	1.000	−.01%	+	0.13
24	Welthandelsorganisation	0.516	−6.94%		0
10	Krieg und Radio	0.015	−11.47%		0
29	Erster Nobelpreis für Wirtschaft	0.072	−11.99%		0
39	Investitionen in Osteuropa oder Rußland	0.041	−16.99%		0.15
34	Alkoholkonsum in Europa	0.060	−24.18%		0
3	Drogen in Holland	0.185	−55.66%		0

the concept model decreased the score. For future experiments it might be interesting to test whether a trade-off weighting between translation and concept model conditioned on a reliability score of the machine translation improves performance.

5 Future Work

Our experiments have been done in a vector space retrieval framework, because this made model combination straightforward and allowed direct comparison to other experimental setups reported in the literature. However, it would be interesting to include the cross-lingual LDA-model in a language model setup, similar to [22] in the monolingual case. The inclusion of ESA in such a model would be more complicated. We also leave experiments with more language pairs for future work.

While we have indicators to when concept models are beneficial for word-based retrieval, the effects that take place when combining LDA and ESA would be more difficult to uncover. Research in this direction could focus on the influence of term-weighting in ESA and on that of disambiguation by context in the case of LDA.

6 Conclusion

For cross-language retrieval, it is essential to have a cross-language bridge that is immediately available, in the best case for many languages and with up-to-date vocabulary. To work in practice, a method to extract such bridging information should not rely on specialized algorithms, but on approved techniques for which robust implementations exist that run on large computing facilities. In this work we have shown that Wikipedia is a valuable bridging source and that standard (monolingual) LDA can be applied to multilingual training data when care is taken for suitable length normalization. Thus, we get an improvement of 325% *mrr* compared to a non-competitive score previously reported for LDA with non-normalized Wikipedia data on the Multext corpus.

A second finding is that simple model combinations reliably increase performance. For retrieval of document translations (mate retrieval) the combination of ESA and LDA achieves scores 16% *mrr* better than reported so far for ESA alone. Concept models based on Wikipedia are also complementary to word based retrieval using machine translation output, here we observe an increase by 21% *map* compared to a machine translation base-line.

While ESA performs better than LDA for mate retrieval, this ranking is reversed for the more relevant task of query-based retrieval. This may be because commonly used ESA-parameters have been tuned for retrieval of document translations.

Acknowledgements

We want to thank Matt Lease and the Spoken Language Systems QA-group for helpful discussions.

References

[1] M. Anderka and B. Stein. The ESA retrieval model revisited. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 670–671. ACM, 2009.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] M. Braschler. CLEF 2000-Overview of results. In *Cross-language information retrieval and evaluation: workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000: revised papers*, page 89. Springer Verlag, 2001.

[4] J.G. Carbonell, Y. Yang, R.E. Frederking, R.D. Brown, Y. Geng, and D. Lee. Translingual information retrieval: A comparative evaluation. In *International Joint Conference on Artificial Intelligence*, volume 15, pages 708–715, 1997.

[5] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, and S. Staab. Explicit vs. latent concept models for cross-language information retrieval. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'09)*, 2009.

[6] W. De Smet and M.F. Moens. Cross-language linking of news stories on the web using interlingual topic modelling. In *The 2nd Workshop on Social Web Search and Mining (SWSM2009)*, 2009.

[7] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[8] S.T. Dumais, T.A. Letsche, M.L. Littman, and T.K. Landauer. Automatic cross-language retrieval using latent semantic indexing. *AAAI Spring Symposuim on Cross-Language Text and Speech Retrieval*, pages 115–132, 1997.

[9] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12, 2007.

[10] T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(90001):5228–5235, 2004.

[11] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM New York, NY, USA, 1999.

[12] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

[13] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, 2005.

[14] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Annual meeting-association for computational linguistics*, volume 45, 2007.

[15] David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889, 2009.

[16] X. Ni, J.T. Sun, J. Hu, and Z. Chen. Mining multilingual topics from Wikipedia. In *Proceedings of the 18th international conference on World wide web*, pages 1155–1156. ACM New York, NY, USA, 2009.

[17] M. Potthast, B. Stein, and M. Anderka. A Wikipedia-based multilingual retrieval model. *Lecture Notes in Computer Science*, 4956:522, 2008.

[18] P. Sorg and P. Cimiano. Cross-lingual information retrieval with explicit semantic analysis. In *Working Notes of the Annual CLEF Meeting*, 2008.

[19] P. Sorg and P. Cimiano. An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. In *Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems (NLDB'09)*, 2009.

[20] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, page 427, 2007.

[21] Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, and Edward Y. Chang. Plda: Parallel latent dirichlet allocation for large-scale applications. In *Proc. of 5th International Conference on Algorithmic Aspects in Information and Management*, 2009. Software available at `http://code.google.com/p/plda`.

[22] X. Wei and W.B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM New York, NY, USA, 2006.

[23] SKM Wong, W. Ziarko, and P.C.N. Wong. Generalized vector spaces model in information retrieval. In *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 18–25. ACM New York, NY, USA, 1985.