# Cross-Lingual Document Similarity

Andrej Muhič, Jan Rupnik, Primož Škraba
*A.I. Laboratory, Jožef Stefan Institute*
*Jamova 39, 10000 Ljubljana, Slovenia*
*E-mail: andrej.muhic@ijs.si, jan.rupnik@ijs.si, primoz.skraba@ijs.si*

***Abstract.*** *In this paper we investigated how to compute similarities between documents written in different languages based on a weekly aligned multi-lingual collection of documents. Computing the cross-lingual similarities is based on an aligned set of basis vectors obtained by either latent semantic indexing or the k-means algorithm on an aligned multi-lingual corpus. We evaluated the methods on two data sets: Wikipedia and European Parliament Proceedings Parallel Corpus.*

**Keywords.** cross-lingual, similarity, LSI, k-means, Wikipedia, information retrieval

## 1. Introduction

The globalization process increased the availability of the multi-lingual information sources and the need for automatic cross-lingual processing tools. The prime example of the process is the Wikipedia - while the majority of the pages in 2001 were written in English, the percentage of English articles dropped to 14% by 2012. Our goal is to find language independent representation of the documents without using machine translation tools which may not always be available. Finding a representation which is valid over multiple languages, we can use well-established machine learning tools designed for monolingual text-mining tasks. The representations can be interpreted as multi-lingual topics, which can be used as proxies to compute cross-lingual similarities between documents.

Our work focuses on studying non-probabilistic approaches to analysis of multi-lingual document collections. The methods we compare are described in section 3.

The paper is organized as follows: we first introduce the setting of multilingual data, then describe the algorithms and datasets. We conclude with the evaluation and discussion.

## 2. Multilingual Data

Our data is a collection of documents in multiple languages along with an alignment with correspondences across languages. Individual documents are represented as vectors by using the standard vector space model where each term corresponds to a word or a phrase in a fixed vocabulary. An aligned corpus is represented by a set of matrices $D^J$ defined as $D^J := [d_1^J, \ldots, d_m^J] \in \mathbb{R}^{n_J \times m_J}$ where $m$ is the number of documents in the aligned corpus and $n_J$ is the dictionary size. Let $N = \sum_{J=1}^{M} n_J$, where $M$ is the number of languages. The multi-lingual document $k$ of the aligned corpus is formed as a block vector $d_k = [{d_k^1}^T \ldots {d_k^T}^M]^T \in \mathbb{R}^N$, and documents are indexed so that $d_k^{J_1}$ and $d_k^{J_2}$ is an aligned document pair for all $k, J_1, J_2$.

Our main interest are similarities between documents. We can measure the similarities between documents of the same language by using the cosine similarity. We note that given a transformation function between dictionaries, we could compute similarities between documents in different languages. However, bilingual-dictionaries are not available for many language pairs in the Wikipedia corpus and we will rely on document correspondences to compute document similarities across languages in the present paper. Note that this implies a correspondence between columns in each $D^J$ for all $J$. Translations between dictionaries would give us row-based correspondences, but we will investigate this approach in future work.

Terms in the dictionary are generally not equally important in determining similarity between documents, so we must preprocess the documents. We first use term frequency ($TF$), to prune away infrequent terms. Rather than take a fixed number of top terms in each document we use an adaptive measure.

Let $f(n)$ be a map which returns numbers of terms appearing at least $n$ times. For each language $k$ we find maximal $n$ such that $\max(10^5, N_k) \leq f(n) \leq 6 \cdot 10^5$, where $N_k$ is number of documents in complete dictionary of language $k$. As we do not apply any stemming and lemmatization this ensures that we get well balanced corpus and retain at least of $10^5$ words for languages with rather small Wikipedias but potentially rich vocabularies. Once this pruning step is complete, we further re-weight the remaining terms. A term weight should correspond to the importance of the term for the given corpus. The common weighting scheme is called Term Frequency Inverse Document Frequency (TFIDF) weighting. An Inverse Document Frequency (IDF) weight for the dictionary term $j$ is defined as $w_j = \log(M/DF_j)$, where $DF_j$ is the number of documents in the $M$ document corpus which contain term $j$. A document TFIDF vector is its original vector multiplied element-wise by the weights. The $j$-th element of a document vector is given by $TF_j \log(M/DF_j)$. Finally, we re-normalize each vector to have Euclidean norm equal to 1.

## 3. Algorithms

We compute the low rank approximation of the term-document matrix using two algorithms: $k$-means[3] and cross-lingual latent semantic indexing(CL-LSI)[4].

The $k$-means algorithm is perhaps the most well-known and used clustering algorithm. In order to apply the algorithm we first merge all the term-document matrices into a single matrix by stacking the individual term-document matrices and discarding nonaligned documents.

$$D_{\text{Total}} = \left[ D^{1^T}, D^{2^T}, \cdots, D^{M^T} \right]^T \qquad (1)$$

such that the columns respect the alignment of the documents. Therefore, each document is represented by a long vector indexed by the terms in all languages. These vectors determine the similarity on which the $k$-means is computed. The algorithm outputs a set of centroids that are then separated to form an aligned basis. As a final step the the projected documents are computed as the least squares solutions.

The next method is CL-LSI which is a variant of LSI [6] applied for the retrieval in multilingual environment. The method is based on computing the singular value decomposition of $D_{Total}$. Since the matrix can be large we can use an iterative method like the Lanczos [7] algorithm with reorthogonalization to find the left singular vectors corresponding to the largest singular values. It turns out that Lanczos method converges slowly as the ratio of leading singular eigenvalues is close to one. Moreover Lanczos method is hard to parallelize. Instead we use randomized version of the singular value decomposition (SVD) described in [1] than can be viewed as block Lanczos method. That enables us to use parallelization and can speed up the computation considerably when multiprocessing is available.

At the end we obtain low rank approximation $D_{\text{Total}} \approx U_k \Sigma_k V_k^T$, where $U_k$ are basis vectors of interest and $\Sigma_k$ is truncated diagonal matrix of singular eigenvalues. Each column $u_i$ of $U_k$ consists of block vectors $u_i = \left[ u_i^{1^T} \quad \ldots \quad u_i^{M^T} \right]^T$. We do not normalize each block $j$ as this would destroy the low rank approximation. We obtain the aligned reduced basis $U_{\text{aligned}} = \left[ U^{1^T} \quad \ldots \quad U^{M^T} \right]^T$, where $D_j = U^j \Sigma_k V_k^T$. The reduced language free representation for language $j$ and document $d$ is computed as the weighted least square solution $\Sigma_k^{-1} U^{j^+} d$, where + denotes the pseudo inverse of a matrix that is is used because columns of $U^j$ do not form an orthogonal basis. The numerical implementation of the least squares is done by QR algorithm.

## 4. Data Set

To investigate the empirical performance of the low rank approximations we will test the algorithms on a large-scale, real-world multilingual dataset that we extracted from Wikipedia by using inter-language links as an alignment. This results in a large number of weakly comparable documents in more than 200 languages. Wikipedia is a large source of multilingual data that is especially important for the languages for which no translation tools, multilingual dictionaries as Eurovoc, or strongly aligned multilingual corpora as Europarl are available. Documents in different languages are related with so called 'inter-language' links that can be found on the left of the Wikipedia page. The Wikipedia is constantly growing. There are currently four Wikipedias with more than $10^6$ articles, 40 with more than $10^5$ articles, 100 with more than $10^4$ articles, and 216 with more than 1000 articles. Wikipedia uses special user-friendly markup language that is very easy to write but very hard to parse. Simplicity of language can cause ambiguities and moreover it is constantly changing. For example, separator | can be used in different contexts.

Wikipedia raw xml dumps of all currently 270 active editions were downloaded from the Wikipedia dump page. The xml files are too large to be parsed with DOM like parser that needs to store the whole xml tree in the memory, instead we implemented Sax like parser that tries to simulate behavior of Wikipedia official parser and is as simple, fast and error prone as possible. We parse all Wikipedia markup but do not extend the templates. Each Wikipedia page is embedded in the page tag. First we check if the title of the page consists of any special namespace and do not process such pages. Then we check if this is a redirection page and we store the redirect link as inter-language links can point to redirection link also. If nothing of the above applies we extract the text and parse the Wikipedia markup. Currently all the markup is removed.

We get inter-language link matrix using previously stored redirection links and inter-language links. If inter-language link points to the redirection we replace it with the redirection target link. It turns out that we obtain the matrix $M$ that is not symmetric, consequently the underlying graph is not symmetric. That means that existence of the inter-language link in one way (i.e. English to German) does not guarantee that there is an inter-language link in the reverse direction (German to English). To correct this we transform this matrix to symmetric by computing $M + M^T$ and obtaining an undirected graph. In the rare case that we have multiple links pointing from the document, we pick the first one that we encountered. This matrix enables us to build an alignment across all Wikipedia languages.

The European Parliament Proceedings Parallel Corpus v6 (EuroParl)[8], a corpus released by the EU Parliament. This source offers a large number of comparable documents in multiple languages. In particular, EuroParl (Release v6) provides transcripts of parliamentary session in almost all the EU languages that are professional translations of each other. To create the document, within each file, we create a document for each speech ID[1](so each speech is a document). The documents were aligned by speaker id. The aligned corpus contains approximately 21000 documents in the 21 languages of the EU.

## 5. Evaluation

We measure the performance of the low rank approximation using two metrics: mean average precision mate retrieval and correlation between monolingual similarity profiles between the query document and its nearest neighbour in the common representation space. For each evaluation, we randomly select a training set and test set from the data.

The first evaluation criteria we is the *mean average precision mate retrieval score* (AMPMR). This measures the similarity between the documents and their translations in the common vector space induced by the latent model. Good models map the docu-

[1] In the corpus, this is referred to as the speaker ID.

ments close to their translations - indicating that some language independent (semantic) information was captured. We evaluate each latent model (given by projection operators $P_1$ and $P_2$) by considering a pair of aligned test sets $T_1$ and $T_2$ in languages $L_1$ and $L_2$. We select a query document $q_1 \in T_1$ and denote the corresponding translated document $q_2 \in T_2$. We then compute the projections $P_1 q$ and $P_2 T_2$ and rank the elements of $P_2 T_2$ by their similarity to $P_1 q$ in the projection space (measured by cosine similarity). The mean average precision mate retrieval score is the inverse of the rank of $P_2 q_2$.

This score does not give us a complete picture. The low rank of a mate document does not necessarily indicate poor performance if the documents which outranked it share similar content to the query. Therefore, we compute an alternative performance measure: *correlation between monolingual similarity profiles* (CMSP). As before, we choose a query document, target test corpus, and project them to a common vector space. From the target corpus, we select the closest document $r \in L_2$ to the query in the projection space. We then compute two similarity profile vectors: $v_1$ contains the monolingual cosine similarity between $q_1$ and all the documents in $T_1$ and similarly $v_2$ contains the cosine similarities between $r$ and $T_2$. The score is the correlation coefficient between $v_1$ and $v_2$. Due to space constraints we display only the first evaluation results.

The results for the AMPMR for several fitting parameters are in Tables 1, 2, and 3. We give the result for the pairwise bilingual retrieval over all queries in each test set, all pairs of languages and over ten repetitions of the choice of training and test data. The Wiki top 9 experiment was done on 65000 training documents and retrieval was tested on remaining 5288 documents in the following languages: English, German, French, Dutch, Italian, Polish, Spanish, Russian and Japanese. The second set of languages considered was: English, German, Spanish, Chinese, Slovenian, Catalan and Croatian. The second set contains 13000 aligned documents where 10.000 were used for training. The second set contains the official languages of the X-LIKE European project and will be denoted by Xlike. Figures 1 and 2 illustrate that precisions increases with the size of the basis and number of train documents.

Table 1: **Pairwise retrieval CL-LSI (Xlike)**

| en | de | es | zh | sl | ca | hr |
|---|---|---|---|---|---|---|
| 0 | 0.89 | 0.87 | 0.7 | 0.75 | 0.79 | 0.71 |
| 0.89 | 0 | 0.81 | 0.66 | 0.73 | 0.76 | 0.7 |
| 0.88 | 0.84 | 0 | 0.65 | 0.7 | 0.79 | 0.69 |
| 0.75 | 0.7 | 0.69 | 0 | 0.61 | 0.66 | 0.59 |
| 0.76 | 0.75 | 0.72 | 0.59 | 0 | 0.69 | 0.68 |
| 0.81 | 0.78 | 0.81 | 0.63 | 0.69 | 0 | 0.67 |
| 0.75 | 0.72 | 0.71 | 0.58 | 0.67 | 0.68 | 0 |

Table 2: **Pairwise retrieval CL-LSI (Wiki)**

| en | de | fr | nl | it | pl | es | ru | ja |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.91 | 0.9 | 0.88 | 0.89 | 0.86 | 0.9 | 0.86 | 0.88 |
| 0.92 | 0 | 0.89 | 0.88 | 0.87 | 0.86 | 0.88 | 0.86 | 0.85 |
| 0.9 | 0.9 | 0 | 0.87 | 0.87 | 0.85 | 0.87 | 0.84 | 0.84 |
| 0.9 | 0.88 | 0.86 | 0 | 0.84 | 0.84 | 0.86 | 0.82 | 0.83 |
| 0.89 | 0.87 | 0.87 | 0.84 | 0 | 0.83 | 0.87 | 0.83 | 0.83 |
| 0.87 | 0.87 | 0.85 | 0.84 | 0.83 | 0 | 0.85 | 0.83 | 0.83 |
| 0.91 | 0.89 | 0.87 | 0.85 | 0.87 | 0.85 | 0 | 0.85 | 0.85 |
| 0.88 | 0.87 | 0.84 | 0.82 | 0.84 | 0.84 | 0.85 | 0 | 0.83 |
| 0.89 | 0.87 | 0.85 | 0.83 | 0.83 | 0.83 | 0.85 | 0.84 | 0 |

## 6. Discussion

The results illustrate that LSI outperforms $k$-means in terms of our evaluation criteria. We believe that this is due to LSI capturing the word co-occurrence patterns. The performance of LSI increases with number of training documents and size of the basis. The

Table 3: **Pairwise retrieval K-means (Wiki)**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0 | 0.75 | 0.74 | 0.68 | 0.71 | 0.65 | 0.75 | 0.65 | 0.66 |
| 0.73 | 0 | 0.69 | 0.68 | 0.67 | 0.63 | 0.67 | 0.63 | 0.62 |
| 0.73 | 0.7 | 0 | 0.64 | 0.69 | 0.63 | 0.7 | 0.61 | 0.6 |
| 0.68 | 0.69 | 0.63 | 0 | 0.61 | 0.59 | 0.63 | 0.58 | 0.57 |
| 0.73 | 0.7 | 0.7 | 0.63 | 0 | 0.63 | 0.7 | 0.61 | 0.6 |
| 0.65 | 0.65 | 0.62 | 0.61 | 0.61 | 0 | 0.62 | 0.6 | 0.56 |
| 0.76 | 0.69 | 0.71 | 0.66 | 0.69 | 0.64 | 0 | 0.62 | 0.61 |
| 0.65 | 0.65 | 0.62 | 0.6 | 0.59 | 0.61 | 0.61 | 0 | 0.55 |
| 0.7 | 0.65 | 0.62 | 0.59 | 0.6 | 0.6 | 0.64 | 0.57 | 0 |

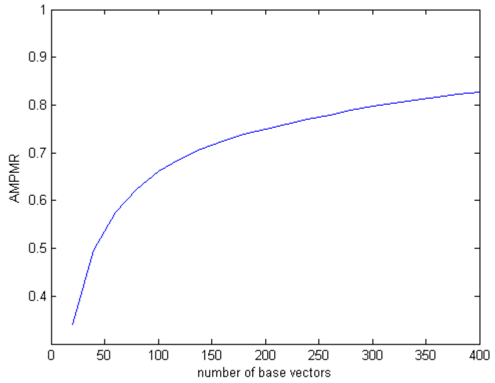

Figure 1: **AMPMR and number of base vectors (Wiki)**

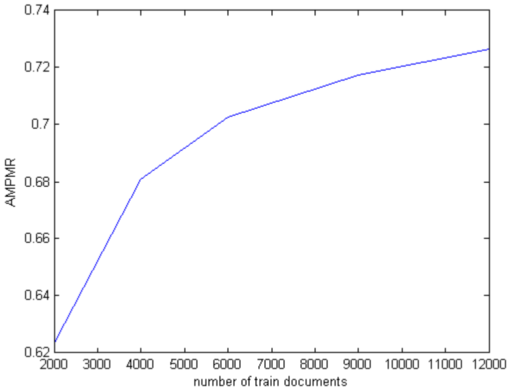

Figure 2: **AMPMR and number of train documents (Xlike)**

weighting done with singular eigenvalues typically increases the performance up to 5%. The AMPRM score rises with the number of available train documents and size of the basis as illustrated in Figures 1 and 2. When forming an aligned corpus, the normalization of document language should be done. Otherwise we always get the rank $k$ approximation that is biased towards the language with documents that have larger number of words. This is undesirable and decreases the quality of the basis.

We also tested CS-LSI on EuroParl corpus and obtained 0.99 AMPMR score. That also indicates that performance of CS-LSI increases with the quality of the corpus. However, additional experimentation is required before we can draw conclusions.

Finally, we note that the results show there is still room for an improvement. We get decent results for small Wikipedias that could be possibly improved by using hub languages. For example two Wikipedias can have almost empty intersection but nevertheless sufficiently large intersection with English Wikipedia. Another option is to allow only partially aligned corpus with empty documents. Preliminary experiments indicate that this leads to decreased performance. Future work includes comparing our method with bilingual dictionary based phrase translations for language pairs with such resources.

## 7. Conclusion

Our work focuses on finding language independent representations of documents written in different languages. The representations are realized as sets of multi-lingual topics that can be used as proxies to compare documents. We experimented with two methods to compute the representations: CL-LSI and k-means. Our experiments indicate that CL-LSI outperforms $k$-means on information retrieval tasks based on the resulting similarity function. Future work includes incorporating bilingual dictionaries and experimenting with probabilistic approaches to topic modelling.

## 8. Learned Concept Vectors

For the illustrative purposes we include the top 20 vectors we learn, each dimension five most significant words for each of the multilingual topics. We only include 2 top Wikipedia languages here. Some words as fontsize:xs are clearly result of parsing error. Otherwise words in aligned English and German top vectors confirm the expected similarity.

```
bar text fontsize:xs 8,5 till he his
b film her cup league club season
team album film band she award film
album d b she calendar onlyinclude
emperor julian year american b d
actor player bar text fontsize:xs
8,5 till he his her she duke album
band comune albums province bar text
```

```
fontsize:xs 8,5 till county comune
album band italy bar text fontsize:xs
8,5 till 1r ret 2r she her 1r bc 2r
bar round flag species party she her
flag ret species university calendar
ret bc prix formula racing flag
bc messier star stars asti turin
piedmont alessandria bc

bar gemeinde text km2 provinz
er deutscher the us of fc saison
nationalmannschaft er league the
film album oder of deutscher the
film amerikanischer us kalender
jahreswechsel a¨ra zeitrechnung
chr deutscher amerikanischer us
politiker komponist deutscher bar
amerikanischer kalender text er
ko¨nig maria war prinzessin album
band stadt single live county album
bar band prozent county prozent album
deutscher maria deutscher stadt insel
fc ko¨nig open hartplatz atp turnier
sand chr open hartplatz qualifiziert
atp flagge arten weibchen partei
ma¨nnchen bar the flagge text of
formel chr dnf rennen v flagge chr
v flaggen sterne bar text till shift
from
```

## 9. Acknowledgements

The authors gratefully acknowledge that the funding for this work was provided by the projecs X-LIKE (ICT-257790-STREP)[9], MultilingualWeb (PSP-2009.5.2 Agr.# 250500)[10], PlanetData (ICT-257641-NoE)[11] and META-NET (ICT-249119-NoE)[12].

## References

[1] Nathan H., Per-Gunnar M., Joel A. T, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, SIAM Rev. 53, 2 (May 2011), 217-288.

[2] *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden.* The Association for Computer Linguistics, 2010.

[3] J. A. Hartigan. *Clustering Algorithms (Probability & Mathematical Statistics).* John Wiley & Sons Inc., 1975.

[4] S.T. Dumais, T.A. Letsche, M.L. Littman, and T.K Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI'97 Spring Symposium Series: CrossLanguage Text and Speech Retrieval*, pages 18–224, 1997.

[5] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58:433–45, 1971.

[6] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407, 1990.

[7] C.F Van Loan G.H. Golub. *Matrix Computations.* Johns Hopkins University Press, 1996.

[8] http://www.statmt.org/europarl/index.html.

[9] http://www.xlike.org/

[10] http://www.multilingualweb.eu/

[11] http://www.planet-data.eu/

[12] http://www.meta-net.eu/

[13] Primoz Škraba, Jan Rupnik, Andrej Muhič. Low-rank approximations for large, multi-lingual data. Avaliable at http://ailab.ijs.si/primoz_skraba/papers/nips_full.pdf.