Multilingual Document Retrieval through Hub Languages

Jan Rupnik, Andrej Muhič, Primož Škraba
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenia

ABSTRACT

In this paper we extend previous work on document retrieval across multilingual corpora. In this setting it is often assumed that we have a certain alignment given based on which we can learn mapping between spaces. In true multilingual corpora however, we often do not have alignments between all languages. There are *hub* languages which have alignments with many other languages. We look at the effectiveness of leveraging these alignments to learn maps which may have small or no alignments given. We test several methods and investigate the performance of various approaches on the Wikipedia dataset.

1 INTRODUCTION

Document retrieval is a well-established problem in data mining. There have been a large number of different approaches put forward in the literature. In this paper we concentrate on a specific setting: multi-lingual corpora. As the availability of multi-lingual documents has exploded in the last few years, the need for automatic cross-lingual processing tools has become apparent. The prime example is Wikipedia - in 2001 the majority of pages were written in English, while by 2012 the percentage of English articles has dropped to 14%. In this context, we look at how to find similar documents across languages. In particular, we do not assume the availability of machine translators, but rather try to frame the problem such that we can use well-established machine learning tools designed for monolingual text-mining tasks.

This work represents the continuation of previous work [4, 2] where we explored representations of documents which were valid over multiple languages. The representations could be interpreted as multi-lingual topics, which were used as proxies to compute cross-lingual similarities between documents. We look at a specific aspect of this problem. The distribution of articles across languages in Wikipedia is not uniform. While the percentage English articles make up as a whole has fallen, in terms of absolute numbers, English is still the largest language. Indeed, there are a number of *hub* languages which have an order of magnitude more articles than other languages.

For document retrieval, if for example, we are looking for a German article comparable to an English article, there is a large alignment between the document corpora (given by the intersection in articles in the two languages) making the problem well-posed. If however we look for a relevant Slovenian article given a Hindi article, the intersection is small, making the problem much harder. However, almost all languages have a large intersection in articles with the *hub* languages, so the question we ask in this paper is: can we exploit hub languages to perform better document retrieval between non-hub languages?

A positive answer would improve cross-lingual analysis in particular between less represented languages. In the following section, we introduce our representation followed by our experiments, which also shed light on the structural properties of the multilingual Wikipedia corpus.

2 DATA MODEL

The key ingredient to our method is a language independent representation upon which we can compute similarities. To model documents we use the standard bag-of-words representation with TF-IDF (term frequency-inverse document frequency) weighting. This representation turns each language into a vector space and cosine similarity induces a metric. From this point on, we operate on languages as metric spaces denoted generically by L. Formally, each document is represented as a point in the metric space. Document comparison can therefore be done by applying the metric on two point $p, q \in L$. This is the starting point of monolingual document retrieval. We now extend this to the multilingual setting. The benefit of this linear representation is that maps between languages (metric spaces) are also linear and we can aim to find a simple (low dimensional/rank) map between the corpora. As input, we are given a partial map in the form of point-

Table 1: Slovenian-Hindi MAPR retrieval using different maps

	sl↦**hi**	hi↦**sl**	sl↦**en**↤hi	sl↦en↦**hi**	hi↦en↦**sl**
LSQ all	0.42	**0.49**	0.38	0.35	0.43
RCCA all	**0.55**	0.45	**0.38**	0.29	0.29
LSQ common	0.42	**0.48**	0.47	0.42	**0.49**
RCCA common	**0.55**	**0.46**	0.39	0.35	0.38
LSQ empty	N/A	N/A	0.27	0.28	**0.35**
RCCA empty	N/A	N/A	**0.32**	0.22	0.21

	sl↔hi	sl↦**en**↤ hi	sl↦en↦**hi**	hi↦en↦**sl**
CL-LSI all	**0.585**	0.35	**0.56**	0.54
CL-LSI common	**0.58**	0.47	**0.61**	**0.61**
CL-LSI empty	0	0.24	**0.48**	0.46

to-point correspondences and we must learn the map (usually through some regression). There are several different formulations of this problem depending on how we *learn* the map addressed in Section 4.2.

In the multilingual setting, for any two languages the number of correspondences may be too small to learn effectively. Therefore, we must go through the hub language (where the number of correspondence with each language may be large) and effectively compose the maps. There are numerous ways to do this which we discuss in the following section.

3 EXPERIMENTS

Experiments were performed using an alignment obtained by Wikipedia on several languages. Specifically we used Slovenian (sl, 91272 words), English (en, 344517 words) and Hindi (hi, 72063 words). The markup in brackets denotes language and number of words in each dictionary. As a preprocessing step, all stub documents with fewer than 20 different words were dropped to improve the quality of the data. The remaining alignment consists of 44426 Slovenian-English correspondences, 4034 Slovenian-Hindi correspondences, 14121 English-Hindi correspondences and 4017 joint Slovenian-English-Hindi correspondences. After stub removal we keep 614 of the initial 1000 test documents and remove them from the training data.

Retrieval can be done in five essentially different ways using our metric space approach:

1. $sl \mapsto \mathbf{hi}$,
2. $hi \mapsto \mathbf{sl}$.
3. $sl \mapsto \mathbf{en} = \mathbf{hub} = \mathbf{en} \leftarrow hi$,
4. $sl \mapsto en \mapsto \mathbf{hi}$
5. $hi \mapsto en \mapsto \mathbf{sl}$.

Bold denotes the space where retrieval is done. The first two represent a direct mapping $sl \leftrightarrow hi$, while the remaining methods map to a common hub space (in this case English), with retrieval occurring either in the target language or the hub language.

To see the amount of information present in the hub languages, we performed tests on three substantially different datasets

- *all* – we use all alignment information available,
- *common* – we use only alignment information consistent through all three languages
- *empty* – we remove all common alignment to simulate the case where we are forced to use hubs.

The evaluation criteria we use is the *mean average precision mate retrieval score* (MAPR). This enables us to compute a similarity between the documents and their translations in the common vector space induced by the latent model or mapping in the common space. Good models will map documents close to their translations – this indicates that some language independent (semantic) information was captured.

Let the individual language we are considering be denoted by L_1 and L_2. Each (latent) model is given by projection operators P_1 and P_2, where one can be identity. We evaluate each model by considering a pair of aligned test sets X and Y in L_1 and L_2. We select a query document $x \in X$ and denote the corresponding translated document $y \in Y$. We then compute the projections P_1x and P_2Y and rank the elements of P_2Y by their similarity to P_1x in the projection space (measured by cosine similarity). The mean average precision mate retrieval score is the inverse of the rank of P_2y. If only one score is displayed, then this is the average of the inverse of the rank of P_1x and the inverse of the rank of P_2y.

3.1 Methods used

In addition to studying the difference in performance depending on which space we perform the retrieval in, there is also the question of how we find the maps.

One approach is to learn the map from the aligned sets X and Y to use a least squares low rank approach.

Table 2: CL-LSI MAPR retrieval in common semantic space

	sl	**en**	**hi**
sl	0	0.77	**0.45**
en	0.73	0	0.64
hi	**0.38**	0.67	0

All

	sl	**en**	**hi**
sl	0	0.81	**0.6**
en	0.77	0	0.71
hi	**0.61**	0.76	0

Common

	sl	**en**	**hi**
sl	0	0.37	**0.22**
en	0.49	0	0.36
hi	**0.11**	0.29	0

Empty

Table 3: CL-LSI MAPR retrieval, full pairwise space

	sl	en	hi
sl	0	0.82	**0.57**
en	0.77	0	0.73
hi	**0.6**	0.78	0

All

	sl	en	hi
sl	0	0.81	**0.58**
en	0.77	0	0.71
hi	**0.58**	0.77	0

Common

	sl	en	hi
sl	0	0.78	0
en	0.7	0	0.71
hi	0	0.77	0

Empty

That is, we find W of rank k with which minimizes $\min ||WX - Y||_F$ where k is an input parameter. The solution can be obtained using a truncated SVD of the input X, $W = YX^+$, $X = U\Sigma V^*$, $X_k^+ = V_k\Sigma_k^{-1}U_k$, where $+$ denotes the pseudo inverse of matrix X. To speed up the computation, a low rank approximation of matrix Y is used. We always use truncated SVDs of size 1000. This approach is denoted as LSQ.

Another method that can be used to relate two aligned sets is regression canonical correlation analysis (RCCA) that is described in [3]. Essentially, this results in the map $q \mapsto (XX')^{-1}XY'q \approx U_k\Sigma_k^{-1}V_k^*Y'q$. Note that this must be used on (implicitly) centered data. Centering explicitly however, is not feasible due to the large number of words which would result in prohibitive RAM requirements. Again we use low rank approximations of implicitly centered Y's to reduce the time complexity and space complexity.

The third method we use is CL-LSI, latent semantic indexing. It is described in [1]. This method enables us to compare documents in the common semantic space. For the sake of clarity, we described this method only for two or three aligned document sets X_1, X_2, X_3. First, we do the SVD decomposition of the glued aligned documents, then we decouple the basis and map in the common subspace.

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \Sigma V^*, \quad \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix} \Sigma V^*,$$

The map to the common semantic space can be described as $x_i \mapsto V\Sigma^+U_i^+x_i$ for $i = 1, 2, 3$, where we overload symbols Σ and V.

To map to English (hub) using the full alignment, we first map Slovenian x_1 to the English word space as $x_1 \mapsto U_2^{12}(U_1^{12})^+x_1$ and Hindi x_3 to English word space as $x_3 \mapsto U_1^{23}(U_2^{23})^+x_3$. This can be done efficiently. Similary we map Slovenian x_1 to the Hindi-English semantic space through English as $x_1 \mapsto U_2^{12}(U_1^{12})^+x_1 = y_1 \mapsto (U_1^{23})^+y_1$ and Hindi x_3 to Hindi-English semantic space $x_3 \mapsto (U_2^{23})^+x_3$.

Mapping through the hub in CL-LSI case enables us to compare documents in the semantic space which as we will see seems to boost performance. In the *all* and *empty* datasets. we glue documents together all three languages despite the lack of an alignment to see how performance degrades in comparison with using the hub.

3.2 Results

As expected, the retrieval is dependent on the mapping used. It is important to note the lack of symmetry in retrieval for the computed RCCA and LSQ mappings. This is to be expected as we only use the information about the target(or alternatively the source) and no common information (as covariance). To illustrate this, RCCA mapping on the *all* dataset, $hi \mapsto \mathbf{sl}$, results in a retrieval score of 0.55, but RCCA mapping other way around on the same dataset, $sl \mapsto \mathbf{hi}$, results in a retrieval score of 0.45.

Better performance could be obtained by using canonical correlation analysis (CCA) although this is a more difficult computationally and has not yet been tested. A similar (dual) situation holds for LSQ mapping where we use only the information about the source space (rather than the target space).

From Table 1 it is not immediately clear which option of using the hub is the best. Using the hub does not improve performance even if we use the whole alignment information available. But is clearly the only option if there is no alignment information through all languages or this alignment is too small.

The CL-LSI method behaves more consistently and outperforms other methods. In this case, the better option is to go through the hub as we can then compare documents in the semantic space at the end. Further tests are

needed to better understand this behavior.

In Tables 2 and 3 we additionally display retrieval using mapping in the common semantic space and using full pairwise alignment, respectively. This gives us an idea about the quality of each mapping.

3.3 Ideal retrieval under misalignment

Consider the *empty* scenario described above: we wish to compare documents between languages L_1 and L_2, but we only have aligned sets for the two languages with a third language L_{hub}. Our aligned sets T_1 and T_2 correspond to $L_1 \mapsto L_{\text{hub}}$ and $L_2 \mapsto L_{\text{hub}}$ respectively.

We assume that no document is shared between T_1 and T_2. Since the alignments are disjoint it may follow that $rank(T_1 \oplus T_2) = rank(T_1) + rank(T_2)$. In such cases no nonzero document can be exactly represented in both bases. Let $f_1 : L_1 \to L_{\text{hub}}$ and $f_2 : L_2 \to L_{\text{hub}}$ represent the regression maps constructed using the alignment. By using the maps we can cast the information retrieval problem between documents in languages L_1 and L_2 as a monolingual information retrieval problem L_{hub}. Since $im(f_1) \subset span(T_1)$ and $im(f_2) \subset span(T_2)$, all inter-lingual similarities will be reduced due to the misalignment of the spaces $span(T_1)$ and $span(T_2)$ (rather than all of L_{hub}). Since the quality of retrieval typically degrades when no direct alignments are available(see Section 4.2), we investigate what is the best possible retrieval under the mis-aligned spaces. That is, what is the highest possible retrieval score on the test set, provided that the images of f_1, f_2 are restricted to $span(T_1)$ and $span(T_2)$ respectively. As in the previous section, the experiment is based on IR between Wikipedia pages written in Slovenian (sl), Hindi(hi) and English (en). The English language represents the hub language with the following document matrices: $T_1 \in \mathbb{R}^{406,044\times 41,529}, T_2 \in \mathbb{R}^{406,044\times 10,331}$, and $T_{\text{test}} \in \mathbb{R}^{406,044\times 604}$.

T_{test} is aligned to A_{test} and B_{test} and $rank([T_1 T_2]) = rank(T_1) + rank(T_2)$. In the ideal case A_{test} and B_{test} would be mapped to T_{test} under f_1 and f_2.

Let $P_X(\cdot)$ denote the orthogonal projection map to the column space of the matrix X. Since the images of f_1 and f_2 are spanned by T_1 and T_2, the test sets would ideally be mapped to $P_{T_1}(T_{\text{test}})$ (ideally projected Slovene test documents) and $P_{T_2}(T_{\text{test}})$ (ideally projected Hindi test documents).

The mean average precision mate retrieval scores we obtain are: 0.995 when A_{test} = query, B_{test} = target and 0.969 when A_{test} = target, B_{test} = query. High retrieval scores indicate that the space of possible maps admits good quality solutions. This result shows potential for improving the retrieval quality.

4 CONCLUSION

The experiments we ran serve two main purposes: the first is an investigation of the performance of using a hub language to enable us to compare languages where alignments may not exist using various different maps and approaches to learning the maps. The second is a structural study, which illustrates how much information is present in the maps. In principle this second part, illustrates that it is possible to find a linear representation in the hub space which yields very high retrieval score, even with no overlap in the alignment. This means that it should be possible to learn maps with very high retrieval rates.

The first set of experiments show that with the approprate preprocessing, going throuh hub languages work reasonably well. However, the lack of symmetry (whereas correspondences are symmetric) in the maps suggests that this may be degrading performance. A method which takes both structures into account may perform better at a higher computational cost. Further, with no alignment there are distribution issues which must be addressed (each language has a different distribution of documents in its monolingual metric space), suggesting that techniques based on transport distance may prove effective.

5 ACKNOWLEDGEMENTS

The authors gratefully acknowledge that the funding for this work was provided by the projects X-LIKE (ICT-257790-STREP), PlanetData (ICT-257641-NoE) and META-NET (ICT-249119-NoE).

References

[1] 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.

[2] Primoz Škraba, Jan Rupnik, Andrej Muhič. Cross-Lingual Document Similarity. In *ITI 2012 Information Technology Interfaces*.

[3] Jan Rupnik, Blaž Fortuna. Regression Canonical Correlation Analysis. Learning from Multiple Sources, NIPS Workshop, 13 Dec 2008 Whistler Canada.

[4] Primoz Škraba, Jan Rupnik, Andrej Muhič. Low-rank approximations for large, multi-lingual data. NIPS Workshop on Low Rank Approximation and Sparse Representation, 2011.