Cross-Lingual Document Retrieval through Hub Languages

Jan Rupnik, Andrej Muhič, Primož Škraba
A.I. Laboratory
Jožef Stefan Institute
Ljubljana, Slovenia
(jan.rupnik),(andrej.muhic),(primoz.skraba)@ijs.si

Abstract

We address the problem of learning similarities between documents written in different languages for language pairs where little or no direct supervision (in the form of a comparable or parallel corpus) is available. To make up for the lack of direct supervision, our approach takes advantage of the fact that they may be linked indirectly by a hub language. That is, correspondences exist between each of the languages and a third, hub language. The main goal of our paper is to explore the viability of cross-lingual learning under such conditions. We propose a method that extracts a set of multilingual topics that facilitate a common representation of documents in different languages. The method is suitable for a comparable multilingual corpus with missing documents. We evaluate the approach in a truly multi-lingual setting, performing document retrieval across eight Wikipedia languages.

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, there is a pressing need for automatic cross-lingual processing tools. The prime example is Wikipedia - in 2001 the majority of pages were written in English, while in 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 [1, 2, 3, 4] where we explored representations of documents which were valid over multiple languages. The representations could be interpreted as multi-lingual topics, which were then 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 of English articles has fallen, English is still the largest language and one of *hub* languages which not only have an order of magnitude more articles than other languages, but also many comparable articles with most of the other languages.

When doing document retrieval, we encounter two quite different situations. 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. Since almost all languages have a large intersection in articles with *hub* languages, 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 vastly improve cross-lingual analysis 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 Approach

We begin by introducing some notation. A *multilingual document* $d = (u_1, \dots u_m)$ is a collection of m documents on the same topic (comparable), where u_i is the document in language i which can be an empty document (missing resource) and d must contain at least two nonempty documents. A comparable corpus $D = d_1, \dots, d_N$ is a collection of multilingual documents. By using the standard vector space model (bag of words) we can represent D as a set of m matrices $X_1, \dots, X_m$, where $X_i \in \mathbb{R}^{n_i \times N}$ is the corpus matrix corresponding to the i-th language, where n_i is the vocabulary size. Let X_i^ℓ denote the ℓ-th column of matrix X_i. A *hub language* is a language with a high proportion of non-empty documents in D, which in case of Wikipedia is English. Let $a(i,j)$ denote the index set of multilingual documents d that contain non-empty u_i and u_j and $a(i)$ denote the index set of multilingual documents that contain u_i.

The goal of our approach is to find a language independent representation of documents by finding a set of mappings $P_1 : \mathbb{R}^{n_1} \to \mathbb{R}^k, \dots, P_m : \mathbb{R}^{n_m} \to \mathbb{R}^k$ that map documents to a common k-dimensional vector space, where similarity between $P_i(x)$ and $P_j(y)$ reflects language independent similarity between x and y.

The first step in our method is to project $X_1, \dots, X_m$ to lower dimensional spaces without destroying the cross-lingual structure. Treating the columns of X_i as observation vectors sampled from an underlying distribution $\mathcal{X}_i \in V_i = \mathbb{R}^{n_i}$, we can analyze the empirical cross-covariance matrices: $C_{i,j} = \frac{1}{|a(i,j)|-1} \sum_{\ell \in a(i,j)} (X_i^\ell - c_i) \cdot (X_j^\ell - c_j)^T$, where $c_i = \frac{1}{a_i} \sum_{\ell \in a(i)} X_i^\ell$. By finding low rank approximations of $C_{i,j}$ we can identify the subspaces of V_i and V_j that are relevant for extracting linear patterns between $\mathcal{X}_i$ and $\mathcal{X}_j$. Let X_1 represent the hub language corpus matrix. The standard approach to finding the subspaces is to perform the singular value decomposition on the full $N \times N$ covariance matrix composed of blocks $C_{i,j}$. If $|a(i,j)|$ is small for many language pairs (as it is in the case of Wikipedia), then many empirical estimates $C_{i,j}$ are unreliable, which can result in overfitting. For this reason we perform the truncated singular value decomposition on the matrix $C = [C_{1,2} \cdots C_{1,m}] \approx USV^T$, where where $U \in \mathbb{R}^{n_1 \times k}, S \in \mathbb{R}^{k \times k}, V \in \mathbb{R}^{(\sum_{i=2}^m n_i) \times k}$. We split the matrix V vertically in blocks with $n_2, \dots, n_M$ rows to obtain a set of matrices V_i: $V = [V_2^T \cdots V_m^T]^T$. Note that columns of U are orthogonal but columns in each V_i are not (columns of V are orthogonal). Let $V_1 := U$. We proceed by reducing the dimensionality of each X_i by setting: $Y_i = V_i^T \cdot X_i$, where $Y_i \in \mathbb{R}^{k \times N}$. The step is similar to Cross-lingual Latent Semantic Indexing (CL-LSI) [5][6] which is less suitable due to a large amount of missing documents. Since singular value decomposition with missing values is a challenging problem and the alternative of replacing missing documents with zero vectors can result in degradation of performance.

The second step involves solving a generalized version of canonical correlation analysis on the matrices Y_i in order to find the mappings P_i. The approach is based on the sum of squares of correlations formulation by Kettenring [7], where we consider only correlations between pairs $(Y_1, Y_i), i > 1$ due to the hub language problem characteristic. Let $D_{i,i} \in \mathbb{R}^{k \times k}$ denote the empirical covariance and $D_{i,j}$ denote the empirical cross-covariance computed based on Y_i and Y_j. We solve the following optimization problem:

$$\underset{w_i \in \mathbb{R}^k}{\text{maximize}} \quad \sum_{i=2}^{m} \left(w_1^T D_{1,i} w_i\right)^2 \quad \text{subject to} \quad w_i^T D_{i,i} w_i = 1, \quad \forall i = 1, \dots, m.$$

By using Lagrangian multiplier techniques (we omit the derivation due to space constraints), the problem can be reformulated as the eigenvalue problem: $\left(\sum_{i=2}^m G_i G_i^T\right) \cdot V = \Lambda \cdot V$, where $G_i = H_1^T D_{1,i} H_i$ and $H_i = Chol(D_{i,i})^{-1}$ where $Chol(\cdot)$ is the Cholesky decomposition: $X = Chol(X)^T Chol(X)$. The vectors w_i can be reconstructed from the dominant eigenvector v, as: $w_1 = H_1 v$ and $w_i \propto H_i G_i^T v$ for $i > 1$ (w_i need to be appropriately normalized). Higher dimensional mappings W_i are obtained in a similar way by setting the constraint $W_i^T D_{i,i} W_i = I$, where I is the identity matrix. The constraint forces the columns of W_i to be uncorrelated (orthogonal with respect to covariance matrix). W_i can be extracted from a set of dominant eigenvectors

of matrix V according to eigenvalues Λ. The technique is related to a generalization of canonical correlation analysis (GCCA) by Carroll[8], where an unknown group configuration variable is defined and objective is to maximize the sum of squared correlation between the group variable and the others. The problem can be reformulated as an eigenvalue problem. The difference lies in the fact that we set the unknown group configuration variable as the hub language, which simplifies the solution. The complexity of our method is $O(k^3)$, whereas solving the GCCA method scales as $O(N^3)$ where N is the number of samples (see [9]). Another issue with GCCA is that it cannot be directly applied to the case of missing documents.

To summarize, we first reduced the dimensionality of our data to k-dimensional features and then found a new representation (via linear transformation) that maximizes directions of linear dependence between the languages. The final projections that enable mappings to a common space are defined as: $P_i(x) = W_i^T V_i^T x$.

3 Experiments

In this section we evaluate our proposed approach to learn the common document representations. The quality of the learned representation is evaluated on the task of cross-lingual document retrieval on Wikipedia. We will first describe the data and then present the evaluation.

To investigate the empirical performance of our algorithm we select a subset of Wikipedia languages containing three major languages, English–*en* (hub language), Spanish–*es*, Russian–*ru*, and five minority (in the sense of Wikipedia sizes) languages, Slovenian–*sl*, Piedmontese–*pms*, Waray-Waray–*war* , Creole–*ht*, and Hindi–*hi*. For preprocessing we remove the documents that contain less than 20 different words (stubs) and remove words occur in less than 50 documents as well as the top 100 most frequent words (in each language separately). We represent the documents as normalized TFIDF[10] weighted vectors.

The evaluation is based on splitting the data into training and test sets (described later). On the training set, we perform the two step procedure to obtain the common document representation as a set of mappings P_i. A test set for each language pair, $test_{i,j} = \{(x_\ell, y_\ell) | \ell = 1 : n(i,j)\}$, consists of comparable document pairs (linked Wikipedia pages), where $n(i,j)$ is the test set size. We evaluate the representation by measuring mate retrieval quality on the test sets: for each ℓ, we rank the projected documents $P_j(y_1), \ldots, P_j(y_{n(i,j)})$ according to their similarity with $P_i(x_\ell)$ and compute the rank of the mate document $r(\ell) = rank(P_j(y_\ell))$. The final retrieval score (between -100 and 100) is computed as: $\frac{100}{n(i,j)} \cdot \sum_{\ell=1}^{n(i,j)} \left(\frac{n(i,j) - r(\ell)}{n(i,j) - 1} - 0.5 \right)$. A score that is less than 0 means that the method performs worse than random retrieval and a score of 100 indicates perfect mate retrieval. The mate retrieval results are included in Table 1.

We observe that the method performs well on all pairs between languages: *en*, *es*, *ru*, *sl*, where at least 50,000 training documents are available. We notice that taking $k = 500$ or $k = 1000$ multilingual topics usually results in similar performance, with some notable exceptions: in the case of (*ht*, *war*) the additional topics result in an increase in performance, as opposed to (*ht*,*pms*) where performance drops, which suggests overfitting. The languages where the method performs poorly are *ht* and *war*, which can be explained by the quality of data (see Table 3 and explanation that follows). In case of *pms*, we demonstrate that solid performance can be achieved for language pairs (*pms*, *sl*) and (*pms*, *hi*), where only 2000 training documents are shared between *pms* and *sl* and no training documents are available between *pms* and *hi*. Also observe that in the case of (*pms*, *ht*) the method still obtains a score of 62, even though training set intersection is zero and *ht* data is corrupted, which we will show in the next paragraph. We now describe the selection of train and test sets. We select the test set documents as all multi-lingual documents with at least one nonempty alignment from the list: (*hi*, *ht*), (*hi*, *pms*), (*war*, *ht*), (*war*, *pms*). This guarantees that we cover all the languages. Moreover this test set is suitable for testing the retrieval thorough the hub as the chosen pairs have empty alignments. The remaining documents are used for training. In Table 2, we display the corresponding sizes of training and test documents for each language pair. The first row represents the size of the training sets used to construct the mappings in low dimensional language independent space using the English–*en* as a hub. The diagonal elements represent number of the unique training documents and test documents in each language.

Table 1: Pairwise retrieval, 500 topics;1000 topics

	en	es	ru	sl	hi	war	ht	pms
en		98 - 98	95 - 97	97 - 98	82 - 84	76 - 74	53 - 55	96 - 97
es	97 - 98		94 - 96	97 - 98	85 - 84	76 - 77	56 - 57	96 - 96
ru	96 - 97	94 - 95		97 - 97	81 - 82	73 - 74	55 - 56	96 - 96
sl	96 - 97	95 - 95	95 - 95		91 - 91	68 - 68	59 - 69	93 - 93
hi	81 - 82	82 - 81	80 - 80	91 - 91		68 - 67	50 - 55	87 - 86
war	68 - 63	71 - 68	72 - 71	68 - 68	66 - 62		28 - 48	24 - 21
ht	52 - 58	63 - 66	66 - 62	61 - 71	44 - 55	16 - 50		62 - 49
pms	95 - 96	96 - 96	94 - 94	93 - 93	85 - 85	23 - 26	66 - 54	

Table 2: Pairwise training:test sizes (in thousands)

	en	es	ru	sl	hi	war	ht	pms
en	671 - 4.64	463 - 4.29	369 - 3.19	50.3 - 2	14.4 - 2.76	8.58 - 2.41	17 - 2.32	16.6 - 2.67
es		463 - 4.29	187 - 2.94	28.2 - 1.96	8.72 - 2.48	6.88 - 2.4	13.2 - 2	13.8 - 2.58
ru			369 - 3.19	29.6 - 1.92	9.16 - 2.68	2.92 - 1.1	3.23 - 2.2	10.2 - 1.29
sl				50.3 - 2	3.83 - 1.65	1.23 - 0.986	0.949 - 1.23	1.85 - 0.988
hi					14.4 - 2.76	0.579 - 0.76	0.0 - 2.08	0.0 - 0.796
war						8.58 - 2.41	0.043 - 0.534	0.0 - 1.97
ht							17 - 2.32	0.0 - 0.355
pms								16.6 - 2.67

We further inspect the properties of the training sets by roughly estimating the fraction `rank(A)/min(size(A))` for each training English matrix and its corresponding mate matrix. Ideally, these two fractions are approximately the same so both aligned spaces should have reasonably similar dimensionality. We display these numbers as pairs in Table 3. It is clear that in the case

Table 3: Dimensionality drift

(en, de)	(en, ru)	(en, sl)	(en, hi)	(en, war)	(en, ht)	(en, pms)
(0.81, 0.89)	(0.8, 0.89)	(0.98, 0.96)	(1, 1)	(0.74, 0.56)	(1, 0.22)	(0.89, 0.38)

of Creole language only at most 22% documents are unique and suitable for the training. Though we removed the stub documents, many of remaining documents are almost the same, as the quality of some minor Wikipedias is low. This was confirmed for Creole, Waray-Waray, and Piedmontese language by manual inspection. The low quality documents correspond to templates about the year, person, town, etc. and contain very few unique words.

We also have a problem with the quality of the test data. For example, if we look at test pair (*war*, *ht*) only 386/534 Waray-Waray test documents are unique but on other side almost all Creole test documents (523/534) are unique. This indicates a poor alignment which leads to poor performance.

4 Conclusions

We proposed a method that enables finding common representations for documents in different languages that is tailored to minority language pairs with limited direct linguistic resources (rare or no comparable document pairs, no dictionary information). We demonstrated that the discovery of cross-lingual mappings is possible even if a pair of languages has no shared linguistic resources, provided that they both share some document correspondences with a hub language.

5 Acknowledgements

The authors gratefully acknowledge that the funding for this work was provided by the projects X-LIKE (ICT-257790-STREP)[11] and META-NET (ICT-249119-NoE)[12].

References

[1] Primoz Skraba, Jan Rupnik, and Andrej Muhic. Low-rank approximations for large, multilingual data. *Low Rank Approximation and Sparse Representation, NIPS 2011 Workshop*, 2011.

[2] Primoz Skraba, Jan Rupnik, and Andrej Muhic. Cross-lingual document similarity. *ITI 2012 Information Technology Interfaces*, 2012.

[3] Primoz Skraba, Jan Rupnik, and Andrej Muhic. Spanning spaces: Learning cross-lingual similarities. *Beyond Mahalanobis: Supervised Large-Scale Learning of Similarity, NIPS 2011 Workshop*, 2011.

[4] Primoz Skraba, Jan Rupnik, and Andrej Muhic. Multilingual document retrieval through hub languages. *Conference on Data Mining and Data Warehouses (SiKDD 2012)*, 2012.

[5] 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.

[6] 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.

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

[8] J. D. Carroll. Generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the American Psychological Association*, pages 227–228, 1968.

[9] Albert Gifi. *Nonlinear Multivariate Analysis*. Wiley Series in Probability and Statistics, 1990.

[10] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. In *INFORMATION PROCESSING AND MANAGEMENT*, pages 513–523, 1988.

[11] `http://www.xlike.org/`.

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