Cross-Lingual Document Analysis

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

Abstract

We present a summary of our work on cross-lingual (CL) document analysis. We focused on learning CL similarity functions and learning language independent document representations. Solutions to either of the two tasks enable us to solve CL text mining tasks (e.g. CL classification, CL retrieval, CL clustering) by using the tools of monolingual text mining. We approached the problems using non-probabilistic methods (typically based on numerical linear algebra) with a special emphasis on scalability. Special attention was devoted to language pairs for which direct training data (translation pairs or comparable documents) was scarce or nonexistent. We showed how to exploit indirect training data through a major (hub) language, such as English.

1 Motivation

As the availability of multi-lingual content on the web 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 in 2012, the percentage of English articles has dropped to 14 %. A standard approach to dealing with multilingual data is by using machine translation techniques. Building good machine translation systems depends on the availability of training data (aligned translations), which is scarce for certain language pairs. However, certain CL tasks admit simpler solutions. In the case of monolingual document retrieval and classification, the widely used vector space model for document representation performs well in practice, even though it discards word order information. Our work relies on the vector space models and is suitable for CL document retrieval, tracking, classification and clustering.

An important aspect of a CL solution is how restrictive its assumptions about the data are. We focus on methods that rely on comparable data (such as the Wikipedia) as opposed to requiring perfect translation pairs and/or bilingual dictionaries. Another important aspect of a CL solution is the computational complexity of training and testing. We focus on methods that can be realized efficiently on a high-end desktop PC.

2 Data representation

To represent a corpus of documents written in the same language we use the vector space model with tf-idf weights [1]. An aligned multilingual collection of documents is represented as a collection of matrices $X_i \in \mathbb{R}^{n_i \times s}$ where i ranges over languages, n_i is the size of the vocabulary of the i-th language and ℓ is the number of documents. The data is aligned over columns (k-th column of X_i and k-th column of X_j represent comparable documents written in languages i and j). Missing documents are represented as zero vectors.

We interpret the collection of aligned documents $\{(X_1(:,k),\ldots,X_m(:,k))\}_{k=1}^{\ell}$ as iid samples obtained from an unknown multivariate distribution over $\mathbb{R}^{\sum_i n_i}$. Therefore our work is based on analyzing empirical covariance matrices $C_{i,j} = \frac{1}{a_{i,j}} X_i X_j^T$, where $a_{i,j}$ is the number of documents that are nonzero simultaneously in languages i and j, where we assume that the data is centered to simplify the presentation.

3 Learning similarity

We will now present our formalism for learning a cross-lingual similarity function [2]. Using the vector space model to represent two documents in languages $x \in \mathbb{R}^{n_i}$ and $y \in \mathbb{R}^{n_j}$, the goal is to define a similarity function for the language pair (i,j):

$$s_{i,j} : \mathbb{R}^{n_i} \times \mathbb{R}^{n_j} \to \mathbb{R}.$$

In our work we focus on finding bilinear similarity functions, i.e.

$$s_{i,j}(x,y) = x^T B_{i,j} y,$$

where $B_{i,j} \in \mathbb{R}^{n_i \times n_j}$. We considered two main approaches to finding B: a regression based approach and a latent space approach.

The **latent space approach** is defined in terms of two linear operators $P_i \in \mathbb{R}^{k \times n_i}$ and $P_j \in \mathbb{R}^{k \times n_j}$ which together with the standard inner product on the *latent space* $Z = \mathbb{R}^k$ define the similarity as:

$$s_{i,j}(x,y) := \langle P_i x, P_j y \rangle = x^T P_i^T P_j y.$$

We considered three choices of latent spaces:

- baseline: $P_i = X_i^T$, $P_j = X_j^T$
- latent semantic analysis [3]
- canonical correlation analysis [4]
- in [5] we included a generalization of CCA to more than two languages
- in [6],[7] we included k-means clustering

The **regression based approach** consists of two ingredients: a linear regression mapping $f : \mathbb{R}^{n_i} \to \mathbb{R}^{n_j}$,

$$f(x) := F \cdot x,$$

where $F \in \mathbb{R}^{n_j \times n_j}$ and an inner product on the target space $\langle \cdot, \cdot \rangle_D$, which are used to define:

$$s_{i,j}(x,y) := \langle f(x), y \rangle_D = f(x)^T D y = x^T F^T D y.$$

Choices of D included the standard inner product I, Mahalanobis inner product $C_{j,j}^{-1}$, which corresponds to whitening, and the covariance matrix $C_{j,j}$ (related to a co-occurrence). We can consider two regression mappings:

- least squares regression (LSQ), where $F = C_{j,i} C_{i,i}^{-1}$
- canonical correlation regression[8] (CCAR), where $F = C_{j,j}^{-1} Cj, i.$

In [2] we elucidated some relationships between the two approaches, for example, CCA is equivalent to the regression based approach with target inner product D as the Mahalanobis inner product. We also conducted an experimental study that compared mono-lingual similarity measures with the proposed cross-lingual similarity measures.

4 Hub languages

Continuing the work started in [7],[9], we looked at a specific aspect of learning CL similarity functions. 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. A large collection (more than $500,000$ articles) of aligned Wikipedia articles is available for the German-English pair, making the learning problem well-posed. If however we consider learning a similarity function for the Slovenian-Hindi pair, there are only a couple of thousand Wikipedia pairs available for learning, which makes the learning problem much harder. However, almost all languages have a large intersection with certain well represented languages, such as English. We refer to English as the hub language. The question we asked in the papers cited is: Can we exploit hub languages to perform better document retrieval between non-hub languages?

In [9] we focused on three languages: Slovenian, Hindi and English (hub). We explored both regression based approaches and common representation based on LSI under various choices of training document collections. We explored several regression scenarios: choosing between LSQ, CCAR and a regression induced by LSI as well as choosing regression sequences (target space and optionally an intermediate space). Here there are five essentially different choices:

1. $sl \mapsto \mathbf{hi}$,
2. $hi \mapsto \mathbf{sl}$.
3. $sl \mapsto \mathbf{en} = \mathbf{hub} = \mathbf{en} \mapsfrom hi$,
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. We showed that direct mappings (e.g. $sl \mapsto \mathbf{hi}$) are more suitable for retrieval when enough data is available, whereas hubs are necessary in the low direct alignment regime. We also noticed that regression based approaches introduced asymmetry in the quality of results as opposed to an approach based on a common representation.

Building on results from [9] we tried incorporating our observations about hub languages in Wikipedia into our methods. We noticed that aligned document sets that do not include English are rare. Taking only cross-covariances that were related to the English language when building common representation models still retains all the vital information. This observation enabled us to formulate a generalization of canonical correlation analysis [10], namely the sum of squared correlations, as an eigenvalue problem which can be solved efficiently. We will now present an overview of the approach.

The goal is to find a language independent representation of documents by finding a set of mappings $P_1, \ldots, P_m$ that map documents to a common k-dimensional vector space. The first step in our method is to project $X_1, \ldots, X_m$ to lower dimensional spaces without destroying the cross-lingual structure. Let X_1 denote the hub language. By using the hub language assumption we compute a singular value decomposition of the stacked cross-covariance matrices, related to the hub language, as:

$$[C_{1,2} \cdots C_{1,m}] = USV^T.$$

We then split the matrix V vertically in blocks with $n_2, \ldots, n_m$ 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) [3],[11], which is less suitable due to a large amount of missing documents. 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 [10], 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:

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

By using Lagrangian multiplier techniques (we omit the derivation due to space constraints), the problem can be reformulated as an eigenvalue problem.

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. To investigate the empirical performance of the proposed method we selected a subset of Wikipedia languages containing three major languages: English (hub language), Spanish, Russian, and five minority (in the sense of Wikipedia sizes) languages: Slovenian, Piedmontese, Waray-Waray, Creole, and Hindi. We evaluated the method on all pairwise retrieval tasks and found that the results were promising; for example, fair retrieval quality was observed for the Hindi-Piedmontesee pair that contained zero direct alignments.

Note on implementation
The preprocessing step requires us to calculate an SVD of a large dense cross-covariance matrix with special structure (fast matrix vector multiplication due to sparse factors). We could use an iterative method like the Lanczos algorithm[12] with reorthogonalization to find the left singular vectors corresponding to the largest singular values. It turns out that the Lanczos method converges slowly as the ratio of leading singular eigenvalues is close to one. Moreover the Lanczos method is hard to parallelize. Instead we use randomized version of the singular value decomposition (SVD) described in [13] than can be viewed as a block Lanczos method. That enables us to use parallelization and can speed up the computation considerably when multiprocessing is available.

5 Conclusions

We presented our work [7],[2],[9],[5] on cross-lingual similarity measures and language independent embeddings. We presented two main approaches to learn document similarities and conducted several experiments on the task of CL information retrieval.

Finding good embeddings in the presence of missing data presents a challenge - for a learning resource, such as the Wikipedia, the topic coverage may vary greatly across languages. This means that certain topics might be difficult to represent in all languages. Our work indicates that learning is possible even for language pairs with no direct alignment information, by leveraging their alignment information with a hub language, such as English.

6 Acknowledgements

The authors gratefully acknowledge that the funding for this work was provided by the projects X-LIKE (ICT-257790-STREP)[14], MultilingualWeb (PSP-2009.5.2 Agr.# 250500)[15], TransLectures (FP7-ICT-2011-7)[16], PlanetData (ICT-257641-NoE)[17], RENDER (ICT-257790-STREP)[18], and META-NET (ICT-249119-NoE)[19].

References

[1] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing And Management*, pages 513–523, 1988.

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

[3] 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.

[4] David R. Hardoon, Sandor Szedmak, Or Szedmak, and John Shawe-taylor. Canonical correlation analysis; an overview with application to learning methods. Technical report, 2007.

[5] Jan Rupnik, Andrej Muhic, and Primoz Skraba. Cross-lingual document retrieval through hub languages. *xLiTe: Cross-Lingual Technologies, NIPS 2012 Workshop*, 2012.

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

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

[8] Jan Rupnik and Blaz Fortuna. Regression canonical correlation analysis. *Learning from Multiple Sources, NIPS Workshop, 13 Dec 2008 Whistler Canada*, 2008.

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

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

[11] 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.

[12] Gene H. Golub and Charles F. Van Loan. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[13] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.*, 53(2):217–288, May 2011.

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

[15] `http://www.multilingualweb.eu/`.

[16] `http://translectures.eu/`.

[17] `http://www.planet-data.eu/`.

[18] `http://render-project.eu/`.

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