Spanning Spaces: Learning Cross-Lingual Similarities

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

Abstract

In analyzing multilingual text corpora, we have the practical problem of computing similarities between documents in different languages. Given two documents in different languages, we use monolingual similarity to an aligned set to compute a similarity across languages. We derive several algorithms and show their relationship the choice of similarity function. We also show experimental results illustrating the approach.

1 Introduction

Many machine learning tasks such as clustering, classification, retrieval, etc. rely on computing a metric to compare (data points). In our motivating example, multilingual corpora, we are usually faced with an overwhelming imbalance: some languages have far more documents than others. Exacerbating the problem is that the intersection can be small or non-existent. Therefore, we must use what alignments exist. In this paper, we derive algorithms based on 2 general approaches. We find that one approach is approximated by cross-canonical correlation analysis (CCA) [1] and latent semantic indexing (LSI)[2] approximates another.

In the next section we derive algorithms based on analyzing the covariance structure of the aligned data. We then conclude by briefly discussing experimental results on the Europarl dataset [3] and generalizations and future directions.

2 Approach

We begin with a collection of documents in two languages[1]. The documents are represented via bag-of-words as vector spaces $\mathbb{X}$ and $\mathbb{Y}$. There are several possible choices of monolingual similarity functions. We consider functions of the form: $x'C^{i}_{\mathbb{X}\mathbb{X}}x$ for $i \in \mathbb{Z}$. For $i = 0$, we recover the squared Euclidean distance and for $i = -1$, we get the Mahalanobis distance.

As input, we take a choice of i (choice of inner product) along with an alignment of $\mathbb{X}$ and $\mathbb{Y}$. The problem is to find a mapping which respects both monolingual similarity and measures cross-lingual similarity. There are two pragmatic approaches to finding the mapping. In the first we reduce the problem to monolingual similarity. We find some mapping $f : \mathbb{X} \to \mathbb{Y}$. For a pair of elements $x \in \mathbb{X}$ and $y \in \mathbb{Y}$, we compute their similarity as the similarity of $f(x)$ and y in $\mathbb{Y}$. The second approach is to find an aligned joint basis for $\mathbb{X}$ and $\mathbb{Y}$, which can be interpreted as mapping both elements to a common space and computing similarity in this common representation.

[1] The extension to multiple language is discussed in the Discussion section.

For the latter approach, we use LSI to find an aligned basis and for the former we used least-squares regression. We tested these approaches for 3 choices of similarity: $i = -1, 0, 1$, corresponding to Mahalanobis distance, Euclidean distance, and the metric weighted by the covariance matrix.

The output of all the algorithms is a linear map (a matrix) denoted B, which computes the cross-lingual similarity as $x^T By$ for $x \in \mathbb{X}$ and $y \in \mathbb{Y}$. The simplest choice for B is $C_{\mathbb{X}\mathbb{Y}}$, the cross-covariance matrix.

LSI: To compute B, we take the first d eigenvalue-eigenvector pairs $\left(\mu_i, [{u_{\mathbb{X}}^i}^T, {u_{\mathbb{Y}}^i}^T]^T\right)$ of the cross-covariance matrix, we obtain the approximation of the full covariance matrix:

$$\begin{bmatrix} C_{\mathbb{X}\mathbb{X}} & C_{\mathbb{X}\mathbb{Y}} \\ C_{\mathbb{Y}\mathbb{X}} & C_{\mathbb{Y}\mathbb{Y}} \end{bmatrix} \approx \sum_{i=1}^{d} \mu_i \begin{bmatrix} u_{\mathbb{X}}^i \\ u_{\mathbb{Y}}^i \end{bmatrix} \cdot \begin{bmatrix} {u_{\mathbb{X}}^i}^T & {u_{\mathbb{Y}}^i}^T \end{bmatrix}, \qquad B = \sum_{i=1}^{d} \mu_i u_{\mathbb{X}}^i {u_{\mathbb{Y}}^i}^T \approx C_{\mathbb{X}\mathbb{Y}} \tag{1}$$

The definition of B comes from comparing the upper-right block. This also shows that LSI approximates the simplest choice of B. It is however far more computationally efficient because it first finds a low-dimensional representation.

Regression: To find B using regression, we compute the optimal regression matrix as $C_{\mathbb{Y}\mathbb{X}}(C_{\mathbb{X}\mathbb{X}})^{-1}$. This corresponds to the mapping f mentioned above. The inner product on the space $\mathbb{Y}$ induces B for computing the similarities between elements in $\mathbb{X}$ and $\mathbb{Y}$: for B it holds that $x^T By = \langle f(x), y \rangle$. Therefore B depends on the metric on $\mathbb{Y}$. For the inner product corresponding to $C_{\mathbb{Y}\mathbb{Y}}$, B can be derived as $C_{\mathbb{X}\mathbb{X}}^{-1} C_{\mathbb{X}\mathbb{Y}} C_{\mathbb{Y}\mathbb{Y}}$.

Note that in the case of regression, we have a choice of which direction we map: $\mathbb{X} \to \mathbb{Y}$ or $\mathbb{Y} \to \mathbb{X}$ except for Mahalanobis metric where the formula for B is symmetric. Due to space constraints we only note here that CCA approximates the least-squares regression for this metric just as LSI approximates the cross-covariance matrix.

3 Experiments

To evaluate a cross-similarity matrix performs, we used monolingual similarity matrices for supervision. All experiments are based on an aligned bilingual corpus of English and German documents and all computations were performed by selecting an aligned subset of 5000 documents T_X, T_Y for learning B and an aligned set of 500 documents R_X, R_Y for testing. For the monolingual similarities $\langle \cdot, \cdot \rangle_X$ and $\langle \cdot, \cdot \rangle_Y$ we compute monolingual similarity scores: $S_X := \langle R_X, R_X \rangle_X$ and $S_Y := \langle R_Y, R_Y \rangle_Y$ and compare the similarities to $S_{XY} := R_X^T B R_Y$. The measure chosen was the average correlation coefficient between rows of S_{XY} and rows of S_Y and the average correlation coefficient between the columns of S_{XY} and rows of S_Y.

We first observe that metrics C_{XX}^i with higher exponent i are easier to learn (the average correlation coefficients between the monolingual and cross-lingual similarity profiles are around 0.9. We also observe that $i = -1$ is the only approach that performs relatively well in the inner product space C_{XX}^{-1} when compared to the other approaches. This is predicted by the theory since the cross-view similarity matrix was derived for that objective. We also note that the ad-hoc approach of $B := C_{XY}$ and its LSI-based approximation perform similarly with only 5 basis vectors used by LSI, which is quite surprising.

			Regression				
	C_{xy}	LSI	$i = -1$	$i = 0$(xy)	$i = 0$(yx)	$i = 1$(xy)	$i = 1$(yx)
C_{xx}^{-1}	0.06	0.05	0.59	0.22	0.21	0.04	0.04
C_{yy}^{-1}	0.06	0.05	0.58	0.22	0.21	0.04	0.04
I_{xx}	0.29	0.28	0.70	0.45	0.45	0.27	0.27
I_{yy}	0.27	0.26	0.66	0.43	0.42	0.25	0.25
C_{xx}	0.94	0.93	0.34	0.87	0.88	0.94	0.94
C_{yy}	0.94	0.93	0.33	0.88	0.88	0.95	0.94

References

[1] H. Hotelling. The most predictable criterion. *Journal of Educational Psychology*, 26:139–142, 1935.

[2] 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.

[3] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*, 2005.

[4] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.