Low-Rank Approximations for Large, Multi-lingual Data

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

Abstract

In this paper we compare low rank approximation methods for data with a particular structure: documents in multiple languages. Rather than looking at only 2 languages in time, we examine the structure in up to 21 languages. The algorithms we choose to compare are k-means, cross-lingual latent semantic indexing(CL-LSI), and multi-view cannonical correlation analysis (mCCA). We test these methods on the European Parliament Proceedings Parallel Corpus.

1 Introduction

When extracting topics from documents in multiple languages, we would like to find topics that are not only important within a language but have an corresponding and equivalent representation in the other languages. This could be considered a language-independent measure of importance. Further, by finding a low-rank approximation which is valid over multiple languages, we can use well-established machine learning tools designed for monolingual tasks. Document collections are a typically represented as high dimensional data sets, making them a prime example for lower dimensional representations.

Our work focuses on studying non-probabilistic approaches to multi-lingual dimensionality reduction. The methods we compare are described in section 3. Recently a probabilistic approach based on latent Dirichlet allocation was proposed [1], however no implementations were readily available. We plan to implement it and include it in subsequent studies.

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 by a vector d indexed by the terms of a dictionary (the i-th element is the term frequency(TF) in the document). For a corpus, we gather document vectors into the term-document matrix D. We choose to index the columns by documents and the rows by terms $D = (d_1, \ldots, d_m)$. Each document is additionally identified by its corresponding language. Each language has its own dictionary independent of other languages.

Primarily, we deal with similarities between documents. Since each language has an independent dictionary, we only define similarity within a language. We note that given a transformation function between dictionaries, we could compute similarities between documents in different languages. However, as translations of individual words is not tailored to the corpus in question and is often not well defined (we often have synonyms, approximate equivalents, etc.), we will use document

correspondences to compute document similarities across languages. Note that this implies a correspondence between columns in each D_k for all k. Translations between dictionaries would give us row-based correspondences, but we will investigate this dimension 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. We find n such that

$$\frac{|f(n+1) - f(n)|}{\text{original \# of terms}} < 0.001 \tag{1}$$

and used these as the terms for the document. Once this pruning step is complete, we further reweight the remaining. 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. A Inverse Document Frequency (IDF) weight for dictionary term j is defined as $w_j = \log(DF_j)$ where DF_j is the number of documents from the corpora 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 $TF_j \log(DF_j)$. 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 three algorithms: k-means[2], cross-lingual latent semantic indexing(CL-LSI)[3], and multi-view canonical correlation analysis (mCCA)[4].

k-means is perhaps the most well-known and used clustering algorithm. The make the data compatible with k-means, we merge all the term-document matrices into a single matrix by stacking the individual term-document matrices.

$$D_{\text{Total}} = \left[D_1^T, D_2^T, \cdots, D_\ell^T \right]^T \tag{2}$$

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. It is these vectors which determine the similarity on which the k-means is computed.

The next method is CL-LSI which is a variant of LSI [5] for more than one view. Each view in this context represents a language. It merges the individual term-document matrices in the same way as we did for k-means. LSI computes a singular value decomposition of D_{Total}. Since the matrix can be large we can use an iterative method like the Lanczos [6] algorithm to find the left singular vectors corresponding to the largest singular values.

Finally, we test mCCA is specifically designed to consider data from multiple sources (in this case languages). Therefore, we do not merge the individual term-document matrices into one as in the other two methods. For each language (view), we estimate the pairwise correlation coefficients, then we try to find vectors which maximize the sum of all pairwise correlations over all languages. This can be written as the following optimization

$$\max_{w_1,\ldots,w_\ell} \sum_{i=1}^{\ell} \sum_{j=i+1}^{\ell} \frac{w_i^T D_i D_j^T w_j}{\sqrt{w_i^T D_i D_i^T w_i} \sqrt{w_j^T D_j D_j^T w_j}} \tag{3}$$

However, this allows only for a 1-dimensional representation (i.e. one vector per language). Since we would like to allow for more vectors per language, we denote the double sum in Equation 3 by mCCA$(w_1, \ldots, w_\ell)$. If corr denotes the correlation, to find M vectors, the optimization objective function becomes

$$\max_{\substack{w_i^{(j)} \; ; \; i=1,\ldots,\ell \\ j=1,\ldots,M}} \sum_{s=1}^{\mathbf{M}} \text{mCCA}\left(w_1^{(s)}, \ldots, w_\ell^{(s)}\right), \qquad \text{s.t.} \qquad \text{corr}\left(w_i^{(s)}, w_i^{(t)}\right) = 0 \quad \forall s \neq t \tag{4}$$

We require that the set of M vectors we return for each language are uncorrelated (so that we do not get copies of a vector) which together maximize the pairwise correlation between languages.

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: the European Parliament Proceedings Parallel Corpus v6 (EuroParl)[7], 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. This can be considered a *gold standard* in terms of multi-lingual data: these documents are professional translations of each other. In the corpus, each day is available as a separate file, although newer data is available in smaller units. To create the document, within each file, we create a document for each speech ID[1](so each speech is a document). In the end, the document contains a document ID, a speech ID, and paragraph markup. The documents were aligned in two ways: by speaker id and by paragraphs. The number of paragraphs was used only if the number was the same for different languages. In total, it contains approximately 21000 documents in the 21 languages of the EU taking up around 2GB od storage.

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 joint representation. 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 documents 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 Euclidean distance). 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.

The results for each of the two measures for several fitting parameters are in Tables 1[2]. We obtain the results by averaging over all queries in each test set, all pairs of languages and over ten repetitions of the choice of training and test data. The experiments are indexed by the number of training samples (Ntrain), the number of languages we consider (NViews) and the dimensionality of the latent space(Ndims). We also ran a preliminary experiment on a data set based on Wikipedia[3] for the same languages. The result is in the final row of Table 1.

6 Discussion

The results illustrate that both LSI and mCCA outperform k-means in terms of our evaluation criteria. We believe that this is due to mCCA and LSI capturing the word co-occurrence patterns, a well-established fact for the LSI method, which also holds for mCCA (The decomposition is based

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

[2] Updated results can be found in the full version of this paper [8]

[3] `http://www.wikipedia.org`

Table 1: Mean Values for Experiments

Parameters	AMPMR			CMSP		
(Ntrain,NViews,Ndims)	k-means	LSI	mCCA	k-means	LSI	mCCA
(100, 3, 5)	0.1096	0.2306	0.2184	0.1135	0.2007	0.1889
(100, 3, 20)	0.3906	0.5292	0.5151	0.3308	0.4466	0.4329
(100, 10, 5)	0.08166	0.1870	0.2006	0.0833	0.1685	0.1725
(100, 10, 20)	0.3204	0.4546	0.4623	0.2686	0.3901	0.3911
(100, 21, 5)	0.09468	0.2040	0.1868	0.1006	0.1789	0.1657
(100, 21, 20)	0.3255	0.4773	0.4577	0.2730	0.4076	0.3878
(1000, 3, 5)	0.1090	0.4059	0.3572	0.1162	0.3268	0.2837
(1000, 3, 20)	0.4180	0.8518	0.8748	0.3465	0.7725	0.7967
(1000, 10, 5)	0.08320	0.3126	0.3801	0.0894	0.2533	0.2965
(1000, 10, 20)	0.3416	0.7695	0.7954	0.2880	0.6921	0.7148
(1000, 21, 5)	0.08427	0.3276	0.3187	0.0783	0.2618	0.2494
(1000, 21, 20)	0.3702	0.8002	0.8075	0.3110	0.7228	0.7240
(1000, 10, 10)	0.2155	0.2335	0.3169	0.2612	0.2779	0.3362

on the inter-lingual covariance matrix). The performance of LSI and mCCA is comparable in all the tested cases. This is unexpected, since LSI discards the correspondence information between feature (word) and language. Note that the above tables are coarse measures, as we average over all pairs of languages. The two algorithms may have different fail cases which may explain why this correspondence information is does not seem important. Furthermore, in our preliminary experiment, mCCA performed significantly better on the Wikipedia data set. However, additional experimentation is required before we can draw conclusions. Finally, we note that the results show there is still room to improve the methods which leads us to believe that a further investigation into low-rank approximations of multilingual structure is warranted.

References

[1] *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.

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

[3] 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.

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

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

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

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

A Learned Concept Vectors

We include the top 20 vectors we learn, each dimension 5 for illustrative purposes. We only include 4 languages here, see [8] for more examples.

closed written statements rule 149	pisemna clanek rozprava ukoncena prohlaseni
women item report voted resolution	zen bodem dalsim pisemne usneseni
you your barroso he mr	jste barroso vam vas mułzete
women gender equality violence men	zen nasili zeny pohlavi zenach
you young economic crisis strategy	jste krize strategie 2020 barroso
vote you minutes details item	hlasovani zapis viz bodem jste
vote minutes details results see	hlasovani zapis viz udajuł pokracujeme
you acp trade agreement european	dohody akt jste partnerstvi komise
euro countries greece acp commission	akt komise eurozony zeme recko
young budget treaty parliament council	parlamentu predsednictvi evropskeho rozpoctu 20
acp agricultural budget agreement farmers	akt dohody zemedelske klimatu zemedelstvi
turkey women market accession policy	zen turecko pristoupeni turecka trhu
presidency trafficking violence council turkey	predsednictvi nasili obchodovani lidmi rady
trafficking vote agricultural policy farmers	obchodovani hlasovani lidmi pisemne soudrznosti
turkey trafficking balkans accession montenegro	makedonie obchodovani turecko lidmi pristoupeni
item israel strategy internal trafficking	bodem strategie obchodovani izrael dalsim
2020 strategy committee eib small	2020 eib bodem podniky demokratuł
turkey acp presidency market internal	akt turecko predsednictvi turecka trh
arctic products greece treaty cosmetic	kosmetickych strategie eurozony arktidy smlouvy
arctic russia next item young	rusko bodem dalsim zdravi arktidy

clos article ecrites 149 debat	schriftliche geschlossen erklarungen aussprache artikel
femmes appelle vote rapport l'ordre	frauen gestimmt bericht schriftlich erklare
vous avez barroso votre pouvez	sie barroso herr antwort selbstverstandlich
femmes hommes l'egalite violence vous	frauen gleichstellung gewalt mannern geschlechter
vous jeunes crise strategie economique	strategie sie 2020 barroso krise
votes vote vous verbal commission	abstimmung kommission the nachster punkt
votes verbal vote proces resultats	abstimmung the protokoll siehe abstimmungsstunde
vous acp turquie europeen l'accord	akp turkei abkommen kommission staaten
euro grece commission acp zone	eurozone griechenland akp kommission lander
jeunes budget traite parlement lisbonne	lissabon parlaments parlament arbeitsmarkt vertrag
acp budget agricole pac l'accord	akp agrarpolitik abkommen versammlung gap
turquie femmes marche interieur politique	turkei frauen binnenmarkt agrarpolitik gap
presidence violence turquie conseil etres	gewalt menschenhandel turkei ratsvorsitz prasidentschaft
traite vote pac etres agricole	menschenhandel agrarpolitik abstimmung gap landwirte
turquie balkans occidentaux montenegro kosovo	turkei menschenhandel mazedonien montenegro kosovo
strategie interieur l'ordre appelle israel	israel binnenmarkt strategie punkt menschenhandel
2020 strategie entreprises bei pme	2020 strategie eib unternehmen kopenhagen
turquie acp jeunes interieur presidence	turkei akp binnenmarkt gewalt versammlung
produits cosmetiques traite zone grece	arktis nanomaterialien eurozone griechenland vertrag
russie arctique l'arctique jeunes sante	russland arktis arktischen punkt frau