Finding Language-Independent Semantic Representation of Text Using Kernel Canonical Correlation Analysis

Alexei Vinokourov ALEXEI@CS.RHUL.AC.UK
John Shawe-Taylor JOHN@CS.RHUL.AC.UK
Computer Science Department, Royal Holloway, University of London, TW20 0EX, UK

Nello Cristianini NELLO@CS.BERKELEY.EDU
BIOwulf Technologies, Berkeley and Division of Computer Science, University of California, Berkeley, US

Abstract

The problem of learning a semantic representation of a text document from data is addressed, in the situation where a corpus of unlabeled paired documents is available, each pair being formed by a short English document and its French translation. This representation can be used either for cross-linguistic retrieval, or, more generally, as a part of a mono-linguistic categorisation or clustering system. By using kernel functions, in this case simple bag-of-words inner products, each part of the corpus is mapped to a high-dimensional space. The correlations between the two spaces are then learnt by using kernel Canonical Correlation Analysis. A set of directions is found in the first and in the second space that are maximally correlated hence forming a semantic representation of the data. Since we assume the two representations are completely independent apart from the semantic content, any correlation between them should reflect some semantic similarity. Certain patterns of English words that relate to a specific meaning should correlate with certain patterns of French words corresponding to the same meaning, across the corpus. Using the semantic representation obtained in this way we report positive results in retrieval of documents, both in a cross language and in single language setting. Our results consistently and significantly outperform those obtained by LSI on the same data.

1. Introduction

Most information retrieval methods depend on exact matches between words in user queries and words in documents. Such methods will, however, fail to retrieve relevant materials that do not share words with users' queries. One reason for this is that the standard retrieval models (e.g. boolean, standard vector, probabilistic) treat words as if they are independent, although it is quite obvious that they are not. A central theme of latent semantic indexing (LSI) (Deerwester et al., 1990) is that term-term interrelationships can be automatically modeled and used to improve retrieval. LSI uses a method from linear algebra, singular value decomposition (SVD) (Golub & van Loan, 1993), to discover the important associative relationships. It is not necessary to use any external dictionaries, thesauri, or knowledge bases to determine these word associations because they are derived from a co-occurence analysis of existing texts. LSI has been adapted to cross-language retrieval (Littman et al., 1998). An initial sample of documents is translated by human or, perhaps, by machine, to create a set of dual-language training documents. After preprocessing documents a common vector-space, including words from both languages, is created and then the training set is analysed in this space using SVD. This method, termed CL-LSI, will be briefly discussed in Section 2.

It has been observed, however, that other statistical and linear algebra methods can provide an improved semantic representation of monolinguistic text data over LSI (Vinokourov & Girolami, 2001)(Vinokourov & Girolami, 2002). In this study we employ kernel Canonical Correlation Analysis (KCCA) (Bach & Jordan, 2001) to learn semantics of text. CCA finds projections in two distinct feature spaces for which the resulting values are highly correlated. Our hypothesis is that finding such correlations between aligned crosslingual corpus will locate the underlying semantics. The directions would carry information about *concepts* which supposedly stood behind the process of generation of the text and, although, they were expressed differently in different languages, they are, nevertheless, semantically the same thing and are presented in the same quantity and quality in both documents of the pairs constituting the dual-language training corpus. We first apply the method to crosslingual information retrieval, comparing performance with a related approach based on latent semantic indexing (LSI) proposed in (Littman et al., 1998). Secondly, we treat the second language as a complex label for the first language document and view the projection obtained by CL-KCCA as a semantic map for use in a multilingual classification task with very encouraging results.

The KCCA machinery will be given in Section 3 and in Section 4 we will show how to apply KCCA to text retrieval with results presented in Section 6.

For additional discussion of kernel methods we refer the reader to (Cristianini & Shawe-Taylor, 2000).

2. Previous work

2.1 CL-LSI

The use of LSI for cross-language retrieval was pioneered by (Littman et al., 1998). LSI uses a method from linear algebra, singular value decomposition, to discover the important associative relationships. An initial sample of documents is translated by human or, perhaps, by machine, to create a set of dual-language training documents $\{x_i\}_{i=1}^N = D_x$ and $D_y = \{y_i\}_{i=1}^N$. After preprocessing documents a common vector-space, including words from both languages, is created and then the training set is analysed in this space using SVD:

$$D = \begin{pmatrix} D_x \\ D_y \end{pmatrix} = U \Sigma V^T, \tag{1}$$

where the i-th column of D corresponds to document i with its first set of coordinates giving the first language features and second set the second language features. To 'translate' a new document (query) q to a language-independent representation one projects (*folds-in*) its expanded vector representation $\tilde{q}$ into the space spanned by the k first eigenvectors U_k:

$$[q] = U_k^T \tilde{q} \tag{2}$$

The similarity between two documents is measured as the inner product between their projections. The documents that are the most similar to the query are considered to be relevant.

2.2 Full eigenvalue decomposition when number of terms is too large

To create a full-dimensional LSI-representation, i.e., when k equals the number of documents N (we assume that $N < M$), one has to use a special technique as one of the dimensions of term-document matrix, the number of terms, is usually too large to perform the SVD of the matrix itself (we have to store the $M \times N$ non-sparse eigenvectors' matrix U in memory). For example, one can take advantage of the so-called 'kernel trick' which helps in cases when the number of data points N is much less than the dimensionality of the data M. Thus, method, discussed in detail in (Cristianini et al., 2002), consists in eigenvalue decomposition of the $N \times N$ kernel matrix $K = D^T D$ instead of $M \times N$ data matrix D:

$$K = V \Sigma^2 V^T = V \Lambda V \tag{3}$$

We hence can find an expression for folding-in new data q independent of the unknown matrix U. Indeed, $[q] = U^T q$, $U = D V \Sigma^{-1}$ and, consequently,

$$[q] = q^T D V \Sigma^{-1} \tag{4}$$

3. Kernel Canonical Correlation Analysis

In this study our aim is to find an appropriate language-independent representation which could assist in multilinguistic information retrieval. Suppose as for CL-LSI we are given *aligned* texts in, for simplicity, two languages, i.e., every text in one language $x_i \in \mathcal{X}$ is a translation of text $y_i \in \mathcal{Y}$ in another language or vice versa. Our hypothesis is that having corpus $\{x_i\}_{i=1}^N$ nonlinearly mapped to $\mathcal{F}_x$ space as $\Phi(x_i)$ and corpus $\{y_i\}_{i=1}^N$ - to $\mathcal{F}_y$ as $\Phi(y_i)$ (with K_x and K_y being respectively the kernels of the two mappings) we can learn (semantic) directions $f_x \in \mathcal{F}_x$ and $f_y \in \mathcal{F}_y$ in those spaces so that projections $(f_x, \Phi(x))$ and $(f_y, \Phi(y))$ of input data images related to different languages onto these directions would maximally correlate with each other and this would mean that these projections could be considered as language-independent representations of the input texts. We have thus intuitively defined a need for the notion of a nonlinear canonical $\mathcal{F}$-correlation $\rho_{\mathcal{F}}$ $(\mathcal{F} = \mathcal{F}_x \times \mathcal{F}_y)$ (Bach & Jordan, 2001) which is defined as

$$\rho_{\mathcal{F}} \quad = \quad \max_{(f_x, f_y) \in \mathcal{F}} \operatorname{corr}((f_x, \Phi(x)), (f_y, \Phi(y)))$$

$$= \max_{(f_x,f_y)\in\mathcal{F}} \frac{\sum_{ij}(f_x,\Phi(x_i))(f_y,\Phi(y_i))}{\sqrt{\sum_i (f_x,\Phi(x_i))^2 \sum_j (f_y,\Phi(y_i))^2}} \quad (5)$$

We search for f_x and f_y in the space spanned by the Φ-images of the data points (*reproducing kernel Hilbert space*, RKHS (Cristianini & Shawe-Taylor, 2000)): $f_x = \sum_l \alpha_l \Phi(x_l)$, $f_y = \sum_m \beta_m \Phi(y_m)$. This rewrites the numerator of (5) as

$$\begin{aligned}
&\sum_{ij}(f_x,\Phi(x_i))(f_y,\Phi(y_i)) \\
=\;& \sum_{ij}\sum_{lm} \alpha_l\beta_m(\Phi(x_l),\Phi(x_i))(\Phi(y_m),\Phi(y_j)) \\
=\;& \alpha^T K_x K_y \beta \qquad (6)
\end{aligned}$$

where α is the vector with components $\{\alpha_l\}$ and β - the vector with components $\{\beta_m\}$. The problem (5) can then be reformulated as

$$\rho_{\mathcal{F}} = \max_{\alpha,\beta} \frac{\alpha^T K_x K_y \beta}{||K_x\alpha||\;||K_y\beta||} \qquad (7)$$

Differentiating the expression under max in (7) with respect to α, taking into account that $\nabla_a||a|| = \frac{a}{||a||}$ and equating the derivative to zero we obtain

$$K_x K_y \beta ||K_x\alpha||^2 - \alpha^T K_x K_y \beta K_x^2 \alpha = 0 \qquad (8)$$

We note that α can be normalised so that $||K_x\alpha|| = 1$. Similar operations for β yield analogous equations that together with (8) can be written in a matrix form:

$$\begin{pmatrix} O & K_yK_x \\ K_xK_y & O \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho \begin{pmatrix} K_x^2 & O \\ O & K_y^2 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \qquad (9)$$

where ρ is the average per point correlation between projections $(f_x,\Phi(x))$ and $(f_y,\Phi(y))$: $\alpha^T K_x K_y \beta$. Equation (9) is known as a generalised eigenvalue problem

$$B\xi = \rho D\xi \qquad (10)$$

where

$$B = \begin{pmatrix} O & K_yK_x \\ K_xK_y & O \end{pmatrix}, \qquad (11)$$

$$D = \begin{pmatrix} K_x^2 & O \\ O & K_y^2 \end{pmatrix},\; \xi = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \qquad (12)$$

The standard approach to the solution of (10) in the case of a symmetric D is to perform incomplete Cholesky decomposition of the matrix D: $D = C^TC$ and define $\varsigma = C\xi$ which allows us, after simple transformations, to rewrite it as a standard eigenvalue problem $C^{-T}BC^{-1}\varsigma = \rho\varsigma$. As D may be singular due to the centering, it needs to be regularised:

$$D = \begin{pmatrix} (K_x + kI)^2 & O \\ O & (K_y + kI)^2 \end{pmatrix} \qquad (13)$$

where the constant k is chosen to be small enough (e.g., 0.001 as advised in (Bach & Jordan, 2001)).

It is easy to see that if α or β changes sign in (9), ρ also changes sign. Thus, the spectrum of the problem (9) has paired positive and negative values between -1 and 1 (see Fig. 1).

As the data has to be centered we can compute the kernel matrix of the centered data (Schölkopf et al., 1999).

4. Cross-linguistic retrieval with KCCA

To fold-in an incoming query q we expand q into the vector representation for its language $\tilde{q}$ and project it onto k canonical $\mathcal{F}$-correlation components:

$$[q] = A^T Z^T \tilde{q} \qquad (14)$$

using the appropriate vector for that language, where A is $N \times k$ matrix where columns are first solutions of (9) sorted by eigenvalue in descending order. Here we assumed that $(\Phi(z),\Phi(\tilde{q}))$ is simply $z^T\tilde{q}$ where Z is the training corpus in the given language:

$$Z = \begin{pmatrix} x_1 & x_2 & \dots & x_N \end{pmatrix} \qquad (15)$$

or

$$Z = \begin{pmatrix} y_1 & y_2 & \dots & y_N \end{pmatrix} \qquad (16)$$

5. Using the semantic space in other applications

The semantic vectors in the given language

$$W = ZA \qquad (17)$$

can be exported and used in some other application, for example, Support Vector Machine classification. We first find common features of the training data used to extract the semantics and the data used to learn SVM classifier, cut the features that are not common and compute the new kernel which is the inner product of the projected data:

$$K'(x_i,x_j) = x_i^T W W^T x_j \qquad (18)$$

The term-term relationship matrix WW^T can be computed only once and stored for further use in the SVM learning process and classification.

6. Experiments

6.1 Mate retrieval

Following (Littman et al., 1998) we conducted a series of experiments with the Hansard collection (Germann,

Table 1. Average accuracy of top-rank (first retrieved) English→French retrieval, %

K	100	200	300	400	FULL
CL-LSI	35±2	56±2	67±3	74±2	84±2
CL-KCCA	76±3	91±1	95±1	96±1	89±10

Table 2. Average precision of English→French retrieval over set of fixed recalls $(0.1, 0.2, \ldots, 0.9)$, %

K	100	200	300	400	FULL
CL-LSI	45±2	51±2	55±2	58±2	64±7
CL-KCCA	67±1	78±1	83±1	86±1	82±9

2001) to measure the ability of CL-LSI and CL-KCCA for any document from a test collection in one language to find its mate in another language. The results are presented in Tables 1 and 2. In all experiments we used the simplest kernel: $k(x_i, x_j) = x_i^T x_j$. The whole collection consists of 1.3 million pairs of aligned text chunks (sentences or smaller fragments) from the 36^{th} Canadian Parliament proceedings. In our experiments we used only the 'house debates' part for which statistics are given in Table 3. As a testing collection we used only 'testing 1'. The raw text was split into sentences with Adwait Ratnaparkhi's `MXTERMINATOR` and the sentences were aligned with I. Dan Melamed's GSA tool (for details on the collection and also for the source see (Germann, 2001)).

The text chunks were split into 'paragraphs' based on '***' delimiters and these 'paragraphs' were treated as separate documents. After removing stop-words in both French and English parts together with rare words (i.e. appearing less than three times) we obtained 5159×12738 term-by-document 'English' matrix and 5611×12738 'French' matrix (we also removed a few documents that appeared to be problematic when split into paragraphs). As these matrices were still too large to perform SVD and KCCA on them, we split the whole collection onto 14 chunks of about 910 documents each and conducted experiments separately with them, measuring the performance of the methods each time on a 917-document test collection. The results were then averaged. Only one - mate document in French was considered as relevant to each of the test English documents which were treated as queries and the relative number of correctly retrieved documents was computed (Table 1) along with average precision over fixed recalls: 0.1, 0.2, ..., 0.9 (Table

Table 3. Statistics for 'House debates' of the 36^{th} Canadian Parliament proceedings corpus.

	SENTENCE PAIRS	ENGLISH WORDS	FRENCH WORDS
TRAINING	948K	14,614K	15,657K
TESTING 1	62K	995K	1067K

2). Very similar results (omitted here) were obtained when French documents were treated as queries and English as test documents. Our results for full CL-LSI are somewhat lower than in (Littman et al., 1998). This is perhaps due to differences in selection of training and test documents. In that work one training set of 982 pairs and one testing set of 1500 pairs were used to evaluate CL-LSI. For some training chunks (for example, for the first one) we observed the performance similar to that reported in (Littman et al., 1998) but for some it was quite different. We have also tested the CL-KCCA method with randomly reshuffled mapping between French and English documents and observed accuracy of about 0.15 which is far lower than learning on the non-random original data. It is worth noting that CL-KCCA behaves differently from CL-LSI over the full scale of the spectrum. When CL-LSI only increases its performance with more eigenvectors taken from the lower part of spectrum (which is, somewhat unexpectedly, quite different from its behaviour in the monolinguistic setting), CL-KCCA's performance, on the contrary, tends to deteriorate with the dimensionality of the semantic subspace approaching the dimensionality of the input data space.

The partial Singular Value Decomposition of the matrices was done using Matlab's 'svds' function and full SVD was performed using the 'kernel trick' discussed in the previous section and 'svd' function which took about 2 minutes to compute on Linux Pentium III 1GHz system for a selection of 1000 documents. The Matlab implementation of KCCA using the same function, 'svd', which solves the generalised eigenvalue problem through Cholesky incomplete decomposition, took about 8 minutes to compute on the same data.

6.2 Pseudo query test

To perform a more realistic test we generated short queries, which are most likely to occur in search engines, that consisted of the 5 most probable words from each test document. The relevant documents were the test documents themselves in monolinguistic retrieval (English query - English document) and their mates

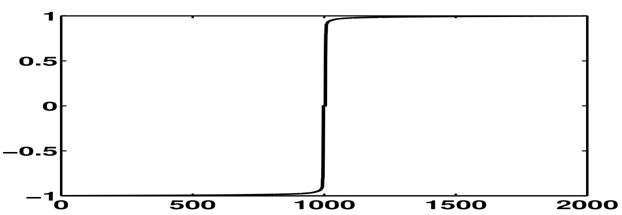

Figure 1. Spectrum of correlation ρ for 1000 English - 1000 French documents corpus.

in the cross-linguistic (English query - French document) test. Table 4 shows the relative number of correctly retrieved as top-ranked English documents for English queries and Table 5 shows the relative number of correctly retrieved documents in the top ten ranked. Tables 6 and 7 provide analogous results but for cross-linguistic retrieval.

Table 4. English-English top-ranked retrieval accuracy, %

K	100	200	300	400	FULL
CL-LSI	17±1	24±1	28±1	31±1	40±3
CL-KCCA	40±2	55±2	61±1	64±1	60±6

Table 5. English-English top-ten retrieval accuracy, %

K	100	200	300	400	FULL
CL-LSI	39±1	47±1	51±1	54±1	63±4
CL-KCCA	83±1	91±1	94±1	94±1	88±5

6.3 Text categorisation using semantics learned on a completely different corpus

The semantics (300 vectors) extracted from the Canadian Parliament corpus (`Hansard`) was used in Support Vector Machine (SVM) text classification (Cristianini & Shawe-Taylor, 2000) of `Reuters-21578` corpus (Joachims, 1998) (Table 8). In this experimental setting the intersection of vector spaces of the Hansards, 5159 English words from the first 1000-French-English-document training chunk, and `Reuters ModApt` split, 9962 words from the 9602 training and 3299 test documents had 1473 words. The extracted 300 KCCA vectors from English and French parts (raw 'KCCA' of Table 8) and 300 eigenvectors from the same data (raw 'CL-LSI') were used in the SVM^{light} (Joachims, 2001) with the kernel (18) to classify the `Reuters-21578` data. The experiments were averaged over 10 runs with 5% each time randomly chosen fraction of training data as the difference between bag-of-words and semantic methods is more contrasting on smaller samples. Comparing 'CL-KCCA' and 'random CL-KCCA' one can conclude that semantics is really extracted using correlations found between translated and original texts. Both CL-KCCA and CL-LSI perform remarkably well when one considers that they are based on just 1473 words. In all cases CL-KCCA outperforms the bag-of-words kernel.

Table 6. English-French top-ranked retrieval accuracy, %

K	100	200	300	400	FULL
CL-LSI	16±1	23±1	27±2	30±1	40±6
CL-KCCA	28±1	37±1	41±1	42±1	33±8

Table 7. English-French top-ten retrieval accuracy, %

K	100	200	300	400	FULL
CL-LSI	47±1	57±2	63±2	66±2	77±10
CL-KCCA	71±2	80±1	82±1	84±1	68±12

Table 8. F_1 value, %, averaged over 10 subsequent runs of SVM classifier with original `Reuters-21578` data ('bag-of-words') and preprocessed using semantics (300 vectors) extracted from the `Canadian Parliament` corpus by various methods. The 5% fraction of `Reuters ModApt` split training data was each time randomly chosen to form training set.

CLASS	EARN	ACQ	GRAIN	CRUDE
BAG-OF-WORDS	81±7	57±3	33±5	13±3
CL-KCCA	**87**±1	**67**±1	59±8	33±7
CL-LSI	77±3	52±3	**64**±14	**40**±2

7. Conclusions

In this work we have applied a kernel version of Canonical Correlation Analysis to the cross-linguistic text retrieval problem. We argue that KCCA model applied to an aligned cross-lingual corpus identifies directions that correspond to the underlying semantics of the documents with the projections into the two languages giving the words most accurately reflecting that semantics. This points a direct route to cross-lingual information retrieval through projecting documents and queries into the common semantic space. Out experiments confirm that CL-KCCA significantly outperforms an earlier analogous method CL-LSI based on Latent Semantic Indexing. Furthemore, we have demonstrated that the semantic space learnt by viewing the French translation as a (complex) label can be used to achieve very good performance on a classification task for data apparently unrelated to the multilingual corpus. This hypothesis is backed by experiments carried out with dual-language proceedings of the Canadian Parliament. As a part of our further study we plan to test KCCA with different kernels. We also note that the approach can be applied with any number of languages and experiments with other languages on such extensive corpora as TREC and CLEF are also part of our future work.

References

Bach, F. R., & Jordan, M. I. (2001). *Kernel independent component analysis* (Technical Report UCB/CSD-01-1166). Division of Computer Science, University of California, Berkley.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.

Cristianini, N., Shawe-Taylor, J., & Lodhi, H. (2002). Latent semantic kernels. *Journal of Intelligent Information Systems*, *18*, 127–152. Special Issue on Automated Text Categorization.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Germann, U. (2001). Aligned Hansards of the 36th Parliament of Canada. http://www.isi.edu/natural-language/download/hansard/. Release 2001-1a.

Golub, G. H., & van Loan, C. F. (1993). *Matrix computations*. The John Hopkins University Press.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. *Proceedings of ECML-98, 10th European Conference on Machine Learning* (pp. 137–142). Chemnitz, DE: Springer Verlag, Heidelberg, DE.

Joachims, T. (2001). SVM^{light} - Support Vector Machine. http://ais.gmd.de/~thorsten/svm_light/.

Littman, M. L., Dumais, S. T., & Landauer, T. K. (1998). Automatic cross-language information retrieval using latent semantic indexing. In G. Grefenstette (Ed.), *Cross language information retrieval*. Kluwer.

Schölkopf, B., Smola, A. J., & Muller, K. (1999). Kernel principal component analysis. In B. Schölkopf, C. J. C. Burges and A. J. Smola (Eds.), *Advances in kernel methods - support vector learning*, 327–352. MIT Press.

Vinokourov, A., & Girolami, M. (2001). Document classification employing the Fisher kernel. *Proceedings of the 23rd BCS-IRSG European Colloquium on IR Research (ECIR'2001)* (pp. 24–40).

Vinokourov, A., & Girolami, M. (2002). A probabilistic framework for the hierarchic organisation and classification of document collections. *Journal of Intelligent Information Systems*, *18*, 153–172. Special Issue on Automated Text Categorization.