EU-IST Project IST-2003-506826 SEKT

SEKT: Semantically Enabled Knowledge Technologies

D1.4.1b Kernel Canonical Correlation Analysis

Blaž Fortuna
Dunja Mladenic, Marko Grobelnik
(Jožef Stefan Institute)

Abstract

Kernel Canonical Correlation Analysis (KCCA) is a technique for finding common semantic features between different views of data. It can be used for finding semantics from different languages that share the same meaning. This information can then be used for mining databases with multilingual text documents.

In this report we present theory behind KCCA and application of it on cross-language information retrieval and classification

Keyword list: Kernel Canonical Correlation Analysis, semantic space, cross-language text mining, language independant representation of text, cross-language information retrieval, cross-language classification

WP1
Prototype/Report PU
Contractual date of delivery: 31.12.2004
Actual date of delivery: 29.12.2004

SEKT Consortium

This document is part of a research pr oject partially funded by the IST Programme of the Commission of the European Communities as project number IST-2003-506826.

British Telecommunications plc.
Orion 5/12, Adastral Park
Ipswich IP5 3RE
UK
Tel: +44 1473 609583, Fax: +44 1473 609832
Contact person: John Davies
E-mail: john.nj.davies@bt.com

Empolis GmbH
Europaallee 10
67657 Kaiserslautern
Germany
Tel: +49 631 303 5540
Fax: +49 631 303 5507
Contact person: Ralph Traphöner
E-mail: ralph.traphoener@empolis.com

Jozef Stefan Institute
Jamova 39
1000 Ljubljana
Slovenia
Tel: +386 1 4773 778, Fax: +386 1 4251 038
Contact person: Marko Grobelnik
E-mail: marko.grobelnik@ijs.si

University of Karlsruhe, Institute AIFB
Englerstr. 28
D-76128 Karlsruhe
Germany
Tel: +49 721 608 6592
Fax: +49 721 608 6580
Contact person: York Sure
E-mail: sure@aifb.uni-karlsruhe.de

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
UK
Tel: +44 114 222 1891
Fax: +44 114 222 1810
Contact person: Hamish Cunningham
E-mail: hamish@dcs.shef.ac.uk

University of Innsbruck
Institute of Computer Science
Techikerstraße 13
6020 Innsbruck
Austria
Tel: +43 512 507 6475
Fax: +43 512 507 9872
Contact person: Jos de Bruijn
E-mail: jos.de-bruijn@deri.ie

Intelligent Software Components S.A.
Pedro de Valdivia, 10
28006
Madrid
Spain
Tel: +34 913 349 797
Fax: +49 34 913 349 799
Contact person: Richard Benjamins
E-mail: rbenjamins@isoco.com

Kea-pro GmbH
Tal
6464 Springen
Switzerland
Tel: +41 41 879 00
Fax: 41 41 879 00 13
Contact person: Tom Bösser
E-mail: tb@keapro.net

Ontoprise GmbH
Amalienbadstr. 36
76227 Karlsruhe
Germany
Tel: +49 721 50980912
Fax: +49 721 50980911
Contact person: Hans-Peter Schnurr
E-mail: schnurr@ontoprise.de

Sirma AI EAD, Ontotext Lab
135 Tsarigradsko Shose
Sofia 1784
Bulgaria
Tel: +359 2 9768 303, Fax: +359 2 9768 311
Contact person: Atanas Kiryakov
E-mail: naso@sirma.bg

Vrije Universiteit Amsterdam (VUA)
Department of Computer Sciences
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands
Tel: +31 20 444 7731, Fax: +31 84 221 4294
Contact person: Frank van Harmelen
E-mail: frank.van.harmelen@cs.vu.nl

Universitat Autonoma de Barcelona
Edifici B, Campus de la UAB
08193 Bellaterra (Cerdanyola del Vall` es)
Barcelona
Spain
Tel: +34 93 581 22 35, Fax: +34 93 581 29 88
Contact person: Pompeu Casanovas Romeu
E-mail: pompeu.casanovas@uab.es

Executive Summary

Kernel Canonical Correlation Analysis (KCCA) is a technique for finding common semantic features between different views of data. It can be used for finding semantics from different languages that share the same meaning. This information can than be used for mining databases with multilingual text documents.

In this report we present a technique for constructing "language independent" representation of text documents. It can be used for cross-language text mining like cross-language information retrieval and cross-language classification. Experimenal results are also presented which show that this approach is promising. The devloped software is also described providing its architecture and users guide. The develoepd software consosts of two utilities, (1) the utility that learns a language independent semantic space for two languages from paired corpus and, (2) the utility that projects documents to the semantic space provided by the firts utility. Both utilities are integrated into our TextGarden library.

Contents

1. Introduction

Canonical Correlation Analysis (CCA) is a method of correlating two multidimensional variables. It makes use of two different views of the same semantic object (e.g. the same text document written in two different languages) to extract representation of the semantic. Input to CCA is a paired dataset *S = {(u$_i$,v$_i$); u$_i$ Î U, v$_i$ Î V}*, where *U* and *V* are two different views on the data – each pair contains two views of the same document. The goal of CCA is to find the common semantic space *W* and the mappings from each *U* and *V* into *W* space. All documents from *U* and *V* can be mapped into *W* to obtain a view independent representation.

Example: Let space *V* be vectors-space model for English and *U* vector-space model for French text documents. Paired dataset is then a set with pairs made of English documents, together with their French translation. The output of CCA on this dataset is a semantic space where each dimension shares similar English and French meaning. By mapping English or French documents into this space, language independent representations are obtained. In this way, standard machine learning algorithms can be used on multi-lingual datasets.

2. Theoretical Foundations

Canonical Correlation Analysis ([1], [2]) can be seen as the problem of finding basis vectors for two sets of variables such that the correlations between the projections of the variables onto these basis vectors are mutually maximized. Canonical Correlation Analysis seeks a pair of linear transformations, one for each of the sets of variables, such that, when the set of variables are transformed, the corresponding co-ordinates are maximally correlated.

Let *S = {(u$_i$,v$_i$); u$_i$ Î U, v$_i$ Î V}* be a paired dataset. By using the CCA, we can find directions *f$_u$Î U* and *f$_v$Î V* in the two spaces so that the projections

{(f$_u$ u$_i$), i = 1, …, N} and *{(f$_v$ v$_i$) where i = 1, …, N*

of the feature vectors of documents from the two views would be maximally correlated. Formally, the CCA is to maximize canonical correlation *r* in space *U ´ V* that is defined as

$$r = \max_{(f_u,f_v)\in U\times V} \frac{\sum_{i=1}^{N} < f_u, u_i >< f_v, v_i >}{\sqrt{\sum_i < f_u, u_i >^2 \sum_i < f_v, v_i >^2}}.$$

In an attempt to increase the flexibility of the feature selection, kernelisation of CCA (KCCA) can be applied to map the hypothesis to a higher-dimensional feature space. There we search for *f$_u$* and *f$_v$* in the space spanned by the corresponding feature vectors, i.e. *f$_u$=å$_l$ a$_l$ u$_l$* and *f$_v$=å$_m$ b$_m$ v$_m$*. The upper equation can be rewritten as

$$\sum_i < f_u, u_i >< f_v, v_i > = \sum_i \sum_{l,m} a_l b_m < u_l, u_i >< v_m, v_i > = a^T K_u K_v b\ ,$$

where *a = (a$_1$,…, a$_N$), b = (b$_1$,…, b$_N$)* and *K$_u$* and *K$_v$* are Gram matrixes of *{u$_i$}* and *{v$_i$}*. In order to force non-trivial learning on the correlation, we introduce a regularization parameter to penalize the norms of the associated weights. The problem becomes

$$r = \max_{a,b} \frac{a^T K_u K_v b}{\sqrt{(a^T K_u^2 a + t a^T a)(b^T K_u^2 b + t b^T b)}}.$$

Because regularized equation is not affected by re-scaling of a or b, optimization problem is subject to the two constraints

$$a^T K_u^2 a + t a^T a = 1,$$
$$b^T K_u^2 b + t b^T b = 1.$$

By using corresponding Lagrangian and *Kuhn-Tucker* conditions, we can rewrite the upper optimization problem as a generalized eigenvalue problem

$$\begin{pmatrix} 0 & K_u K_v \\ K_v K_u & 0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = l \begin{pmatrix} K_u^2 + tI & 0 \\ 0 & K_v^2 + tI \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}.$$

Because of regularization obtained vectors a and b are not equally scaled. This can be solved by normalizing obtained directions. In the upper derivation, we assumed that we have two different views of documents (*U* and *V*). CCA can be generalized to more views, but then the trick to reduce the size of eigen problem cannot be used.
Note that the size of generalized eigen problem is *2N*, where *N* is the size of the paired dataset. This can be reduced by using *incomplete Cholesky decomposition* to *N* or even less when seeking only approximate solution. Algorithms for solving this optimization problem are all of order $O(N^3)$ or less and can be efficiently implemented. For example for *N = 1000* it takes around one minute to solve optimization problem on normal desktop computer.

3. Applications of KCCA

3.1. Labels

A similar problem to CCA is to select features of highest correlation between documents and their labels. The method for finding these features is called Partial Least Squares (PLS) [1]. PLS could also be thought as a method which looks for directions that are good at distinguishing the different labels. Similarity between this problem and CCA can be noticed when viewing labels as another "different view of documents".

3.2. Cross-Language Text Mining

With KCCA we can construct a semantic space into which text documents, written in different languages, can be mapped to obtain language independent representation. This highly reduces the complexity of dealing with different languages since we can apply standard machine learning algorithms to the data mapped into the semantic space. Another method for dealing with multi-lingual datasets is CL-LSI [4].

3.2.1. Text document retrieval

The semantic space for languages can be used at searching databases with documents in different languages. First, all documents from the database are mapped into the semantic space. Than, queries can be viewed as documents and can be mapped into the semantic space. The result of a query is a set of documents from the database that are the closest to the mapped query in the semantic space. The advantage of this approach is that the results are independent of the language in which the query was issued.

This approach was shown and tested in [3] on “house debates” part of 36th Canadian Parliament proceedings corpus. Text chunks were split into paragraphs and paragraphs were treated as separate documents. Part of this dataset was used for generating the semantic space with KCCA and the rest of the documents were used for testing. Short queries were generated from the five most probable words from each test document. The relevant documents were the test documents themselves in monolinguistic retrieval (English query - English document, table 1) and their mates in cross-linguistic (English query - French document, table 2) test. Each test was done for different dimensions *d* of the generated semantic space.

d	100	200	300	400	full
cl-lsi	53	60	64	66	70
kcca	60	63	70	71	73
cl-lsi	82	86	88	89	91
kcca	90	93	94	95	95

Table 1: English to English top-ranked (left) and top-ten (right) retrieval accuracy

d	100	200	300	400	full
cl-lsi	30	38	42	45	49
kcca	68	75	78	79	81
cl-lsi	67	75	79	81	84
kcca	94	96	97	98	98

Table 2 English to French top-ranked (left) and top-ten (right) retrieval accuracy

3.2.2. Text categorization

Another application of the semantic space is categorization of multi-lingual documents. First, the semantic space is generated from the paired dataset with KCCA. Then, the labeled training set for categorization is mapped into the semantic space. Note that these labeled documents do not need to be paired anymore. Even more, they can even come from only one language. Once training set is mapped into semantic space standard classification algorithms can be used, e.g. SVM. Another way of using SVM is to learn classifier on labeled documents from one language and than transfer it trough semantic space into other language's vector-space model.

This approach was shown and tested in [5] on NTCIR-3 patent retrieval test collection, with paired documents in English and Japanese. The classifier was learned on documents in one language and was used to classify documents in another language. The training set for Topic 01 had 827 annotated documents with 26 relevant documents; Topic 07 had 366 annotated documents with 102 relevant documents. The classifier was trained on English training set. Results are in table 3.

d	50	100	150	full	50	100	150	full
Eng-tr	78.1	97.7	99.2	100.0	87.6	93.9	95.8	97.1
Eng-ts	36.0	41.0	44.4	46.9	85.1	87.4	87.0	87.9
Jp-tr	79.4	92.5	98.4	99.2	87.4	92.9	95.4	96.8
Jp-ts	41.1	42.4	48.9	49.1	77.2	77.7	77.3	78.4

Table 3: Average precision [%]: the classifier learned on English training set was used on English training and test sets and on Japanese training and test sets. On left are results for Topic 01 and on right for Topic 07.

3.2.3. Machine Translation and KCCA

The goal of KCCA is to generate language independent semantic space. However, in order to use KCCA, paired dataset is needed. This can be tricked by using machine translation tools, for example *Google Language Tools* (`http://www.google.com/language_tools`) to artificially generate paired dataset from monolinguistic dataset. Semantic space obtained from this kind of paired dataset can than be used for text as described upper.

We are currently doing experiments with this approach and the results achieved so far with this approach are very promising. Experiments were done on big Reuter's dataset in English and French language. News articles from both languages are tagged with categories. Experiments were conducted similarly as upper at cross-language text classification. See table 4 and 5 for the results.

[%]	Precision	Recall	F1	Avg. Prec.
Eng-BOW	86	90	88	83
Fr-BOW	93	74	82	80
Eng-Eng	79	90	84	77
Eng-Fr	86	79	77	73
Fr-Eng	68	90	83	76
Fr-Fr	87	69	77	73

Table 4: Classification of news articles into category CCAT. First two rows show results obtained with normal TFIDF vector representation of articles in original language. Lower rows show results where X-Y means that classifier was trained on language X and tested on language Y, for example Fr-Eng means that it was trained on French and tested on English documents. 10.000 translated documents were used to generate semantic space and 5.000 from this set were used for training. Testing was done over set of 100.000 documents.

[%]	Precision	Recall	F1	Avg. Prec.
Eng-BOW	87	85	86	81
Fr-BOW	95	79	86	81
Eng-Eng	87	76	81	75
Eng-Fr	94	72	81	75
Fr-Eng	86	75	80	75
Fr-Fr	93	74	82	76

Table 4: Classification of news articles into category MCAT.

4. Architecture

Utilities for learning common semantic space for two languages with KCCA are fully integrated into Text Garden. Main utility is named PrSet2SemSpace. It takes as input two Bag-Of-Words files, one for each language, and Paired-Set file with aligned documents from two languages. Documents can also be aligned by paragraphs but that is not necessary. Bag-Of-Words input files define vector-space model for each languages and documents from aligned corpus are transformed into vectors using this models. Output of this utility is a pair of Semantic-Space files (.ssp), one for each document. They define map for documents from vector-space model for each language into common semantic space. See Figure 1 for the diagram of this pipeline. See *Appendix – User Guide* for more details on how to use this utility.

Semantic spaces learned with KCCA can be used by other utilities from Text Garden, that get Semantic-Space file as input. For example ProjBow2SemSpace which takes Bag-Of-Words file and Semantic-Space file as input and gives as output Bag-Of-Words file with projected documents. Output from ProjBow2SemSpace can be use as input to any of algorithems from Text Garden, for example clustering and classification algorithms.

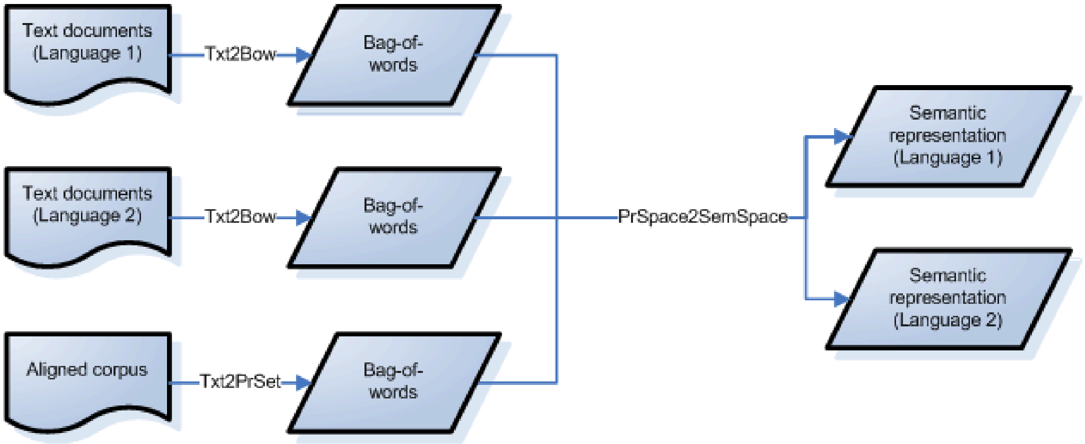
Text documents (Language 1)
Txt2Bow
Bag-of-words
Text documents (Language 2)
Txt2Bow
Bag-of-words
Aligned corpus
Txt2PrSet
Bag-of-words
PrSpace2SemSpace
Semantic representation (Language 1)
Semantic representation (Language 2)

Figure 1. *Pipeline for creating common semantic space for two languages*

5. Future Development

Short term goal is to finish experiments on use of machine translation for generating aligned corpus. We are doing experiments on Reuter's multilingual dataset and Canadian Parliament Proceedings.

We also have a new aligned corpus of European legislation in Slovenian and English and are planning to do experiments on it. Interesting issues that can be addressed are, how KCCA works on Slovenian language, how does it scale, etc. Also, use of KCCA for information retrieval was not explored in details: how to scale search on larger number of documents, how to do indexing, how to rank results, etc.

There are also many other areas that could gain from use of KCCA that we still have to explore.

6. Conclusion

We presented technique for constructing “language independent” representation of text documents. It can be used for cross-language text mining like cross-language information retrieval and cross-language classification. Some results are also presented which show that this approach is promising.

7. Appendix - User Guide

7.1. Paired-Set-To-Semantic-Space

The utility learns language independent semantic space for two languages from paired corpus (“-ips”). It also uses Bag-Of-Word files for each language (“-ibow1”, “-ibow2”). It outputs two Semantic-Space files, one per language (“-ossp1”, “-ossp2”).

Parameter “-t” is regularization parameter τ from upper derivations. Parameter “-tnrm” determines how basis vectors are normalized after learning (“none” means no normalizing, “one” means normalizing to norm 1 and “eigval” means normalizing to its eigenvalue). Parameter “–tnrm” determines stopping criteria for incomplete Cholesky decomposition. Parameter “-docs” determines number of documents from paired corpus that will be randomly selected (randomizer is initialized with parameter

"-seed"). Parameter "-dim" determines dimension of calculated semantic space. Parameter "-len" determines maximal length of documents used for learning from paired corpus. If documents are split into paragraphs and document is longer than maximal length, than only random subset of paragraphs is used. Parameter "-stat" determines if text file with statistics for each semantic space should be made.

```
usaga: PrSet2SemSpace.exe
    -ips:   Input-PrSet-File-Name (default:'')
    -ibow1: Input-Bow-File-Name-For-First-Language (default:'')
    -ibow2: Input-Bow-File-Name-For-Second-Language (default:'')
    -ossp1: Output-Semantic-Space-File-Name-For-First-Language (default:'')
    -ossp2: Output-Semantic-Space-File-Name-For-Second-Language (default:'')
    -t:     Regularization-Parameter-For-KCCA (default:0.5)
    -tnrm:  Correlation-Normalization-Type (none, one, eigval) (default:'one')
    -eps:   Threshold-For-Partial-Gram-Schmidt (default:0.4)
    -docs:  Number-Of-Documents-For-Training-KCCA (default:1000)
    -dim:   Number-Of-Calculated-Dimensions (default:500)
    -len:   Maximal-Length-Of-Training-Document (-1 for no limit)
    (default:1000)
    -seed:  Seed-For-Randomizer (default:0)
    -stat:  Make-Semantic-Space-Statistics (default:'F')
```

7.2. Project-Bow-Of-Words-To-Semantic-Space

The utility projects documents from Bag-Of-Words files ("-bow") to semantic space given with Semantic-Space file ("-issp") and saves them in Bag-Of-Words file ("-obow").

Parameter "-sspdim" determines the number of dimensions that will be used from semantic space. Parameter "-nrm" determines whether projected documents should be normalized.

```
usage: ProjBow2SemSpace.exe
    -issp:    Input-Semantic-Space-File-Name (default:'')
    -ibow:    Input-Bow-File-Name (default:'')
    -obow:    Output-Projected-Bow-File-Name (default:'')
    -sspdim:  Number-Of-Dimensions-For-Projections (-1 for all) (default:-1)
    -nrm:     Normalize-Projected-Vectors (default:'F')
```

8. Bibliography and references

[1] J. Shawe-Taylor, N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004

[2] D. R. Hardon, S. Szedmark, and J. Shawe-Taylor. *Canonical correlation analysis: an overview with application to learning methods*. Technical Report CSD-TR-03-02, Department of Computer Science, Royal Holloway, University of London, 2003.

[3] A. Vinokourov, J. Shawe-Taylor, and N. Cristianini. *Inferring a semantic representation of text via cross-language correlation analysis*. In Advances of Neural Information Processing Systems 15, 2002.

[4] M. L. Littman, S. T. Dumais, and T. K. Landauer. *Automatic cross-language information retrieval using latent semantic indexing*. In G. Grefenstette, editor, Cross language information retrieval. Kluwer, 1998.

[5] Yaoyong Li and John Shawe-Taylor. *Using KCCA for Japanese-English cross-language information retrieval and classification*