Kernel Canonical Correlation Analysis

(Language Independent Document Representation)

Roland Pihlakas

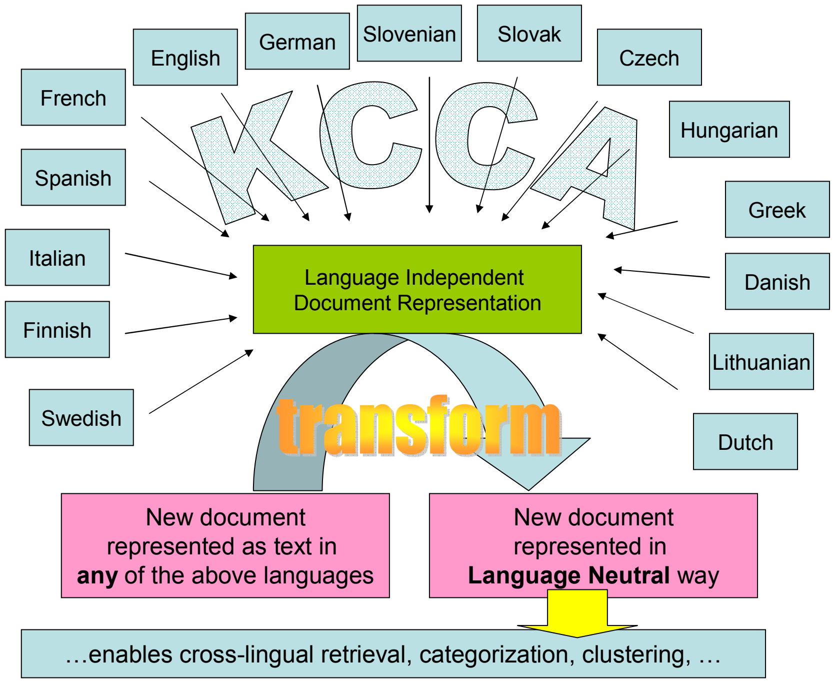
Part of the slides is taken from .PPT with same title by the authors:
Blaz Fortuna, Marko Grobelnik, Dunja Mladenić
Jozef Stefan Institute, Ljubljana

General goal (1/2)

- 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.
- These methods treat words as if they are independent, although it is quite obvious that they are not.

General goal (2/2)

- KCCA enables to represent documents in a “**language neutral way**”
- Intuition behind KCCA:
  1. Given a parallel corpus (such as Acquis)…
  2. …first, we automatically identify language independent semantic concepts from text,
  3. …then, we re-represent documents with the identified concepts,
  4. …finally, we are able to perform cross language statistical operations (such as retrieval, classification, clustering…)
- → It is not neccessary to use any external dictionaries, thesauri, or knowledge bases to determine word associations.
- It can be used both in monolinguistic and multilinguistic retrieval.

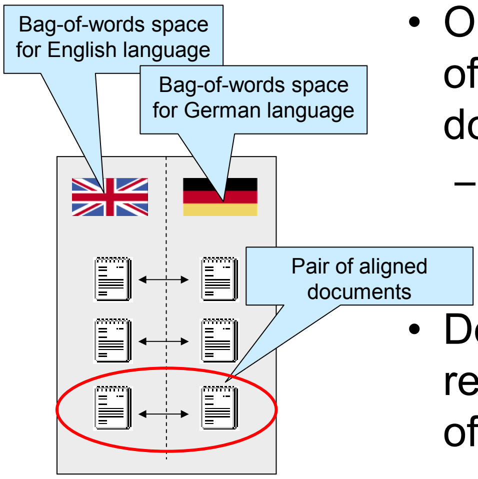


New document represented as text in **any** of the above languages

New document represented in **Language Neutral** way

…enables cross-lingual retrieval, categorization, clustering, …

Input for KCCA



Bag-of-words space for English language

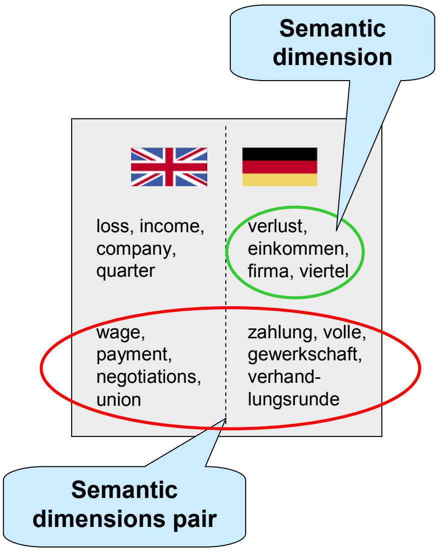
Bag-of-words space for German language

Pair of aligned documents

- On input we have set of aligned documents:
  - For each document we have a version in each language
- Documents are represented as bag-of-words vectors

The Output from KCCA

- **The goal:** find pairs of *semantic dimensions* that co-appear in documents and their translations with high correlation
  - *Semantic dimension* is a weighted set of words.
- These pairs are pairs of vectors, one from e.g. English bag-of-words space and one from German bag-of-words space.



Semantic dimension

loss, income, company, quarter

verlust, einkommen, firma, viertel

wage, payment, negotiations, union

zahlung, volle, gewerkschaft, verhand-lungsrunde

Semantic dimensions pair

What is canonical correlation analysis

- arg $\max_{a,\, b}$ ρ, where ρ = corr(a’X, b’Y)
- X, Y – vectors of random variables
  a, b – vectors we are seeking for
- Typical use for canonical correlation in the psychological context is to take a two sets of variables and see what is common amongst the two sets.
- For example you could take two well established multidimensional personality tests such as the MMPI and the NEO. By seeing how the MMPI factors relate to the NEO factors, you could gain insight into what dimensions were common between the tests and how much variance was shared.

The Algorithm – Theory

- Formally the KCCA solves:
  max(f_x,f_y) corr(<f_x, 🇬🇧 >, <f_y, 🇩🇪 >)
- f_x, f_y – semantic directions for English and German
- (🇬🇧, 🇩🇪) is a pair of aligned documents
- <f_x, 🇬🇧> is presumably the language independent semantic representation.
- f_x is presumably the transformation matrix for specific language to retrieve langage independent representation.
- we search for f_x and f_y such that the resulting representations will be highly correlated.

The Algorithm – Theory

- Formally the KCCA solves:

  max(f_x,f_y) corr$((f_x,$ 🇬🇧 $), (f_y,$ 🇩🇪 $))$
- f_x, f_y – semantic directions for English and German.
- (🇬🇧, 🇩🇪) is a pair of aligned documents.

More details

- "Pair of documents" – in this experiment, is actually a pair of sentences, more specifically, vectors corresponding to bags of the words in the sentences, mapped to high-dimensional space.
- The mapping is to be done using nonlinear function Φ(x), though the Φ was not specified. The aim is to do the Kernel trick.

The Algorithm – Theory

$$\max_{f_x, f_y} corr(< f_x, \Phi(x) >, < f_y, \Phi(y) >)$$

$$\Downarrow$$

$$\boxed{B\xi = \rho D\xi}$$

$$B = \begin{pmatrix} O & K_x K_y \\ K_x K_y & O \end{pmatrix} \quad D = \begin{pmatrix} K_x^2 & O \\ O & K_y^2 \end{pmatrix} \quad \xi = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

$$\Downarrow$$

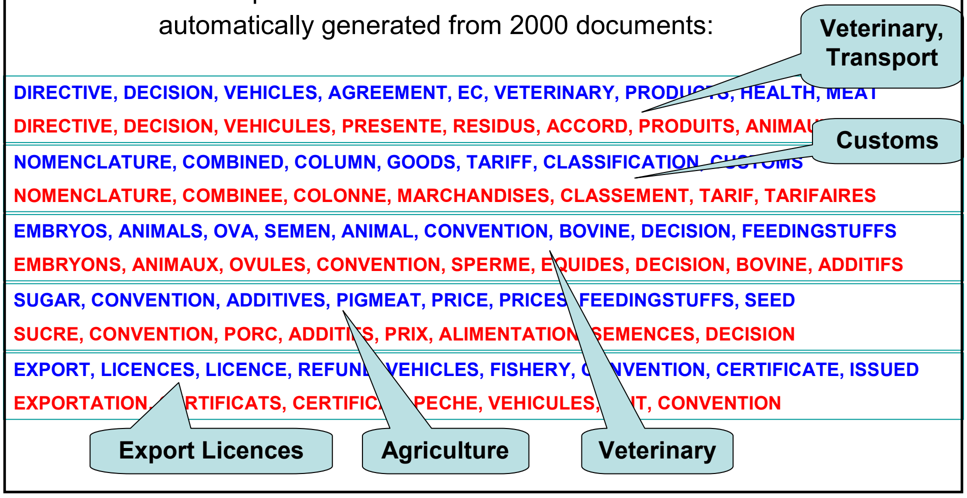
$$f_x = \sum_i \alpha_i \Phi(x_i) \qquad f_y = \sum_j \beta_j \Phi(y_j)$$

The Algorithm – Theory

- K_x , K_y – kernels of the two nonlinear mappings Φ(x) and Φ(y). In the experiment, they actually used simple linear kernel $k(x_i, x_j) = x_i^T x_j$. They plan to use different kernels in the future.
- x, y – corpuses in some specific language consisting of set of documents $x = \{x_i\}^N_{i=1}$
- x_i,y_i – pair of translated texts

Examples of Semantic Dimensions from Acquis corpus: English-French

Most important words from semantic dimensions automatically generated from 2000 documents:



DIRECTIVE, DECISION, VEHICLES, AGREEMENT, EC, VETERINARY, PRODUCTS, HEALTH, MEAT

DIRECTIVE, DECISION, VEHICULES, PRESENTE, RESIDUS, ACCORD, PRODUITS, ANIMAU

NOMENCLATURE, COMBINED, COLUMN, GOODS, TARIFF, CLASSIFICATION, CUSTOMS

NOMENCLATURE, COMBINEE, COLONNE, MARCHANDISES, CLASSEMENT, TARIF, TARIFAIRES

EMBRYOS, ANIMALS, OVA, SEMEN, ANIMAL, CONVENTION, BOVINE, DECISION, FEEDINGSTUFFS

EMBRYONS, ANIMAUX, OVULES, CONVENTION, SPERME, EQUIDES, DECISION, BOVINE, ADDITIFS

SUGAR, CONVENTION, ADDITIVES, PIGMEAT, PRICE, PRICES, FEEDINGSTUFFS, SEED

SUCRE, CONVENTION, PORC, ADDITIFS, PRIX, ALIMENTATION, SEMENCES, DECISION

EXPORT, LICENCES, LICENCE, REFUN VEHICLES, FISHERY, CONVENTION, CERTIFICATE, ISSUED

EXPORTATION, RTIFICATS, CERTIFIC PECHE, VEHICULES, IT, CONVENTION

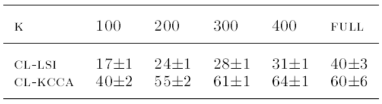
The experiment

- 1.3 million pairs of aligned text chunks (sentences or smaller fragments). Text was taken from 36th Canadian Parliament proceedings.
- But: they removed some words and a few documents that appeared to be problematic when split into paragraphs.
- The result was 5159 x 12738 term-by-"document" English matrix and
  5611 x 12738 term-by-"document" French matrix.
- These matrixes were still too large to perform matrix decompositions, so the experiment was done on 14 chunks of the above data and the results were averaged.
- Computation in Matlab took ~2-8 minutes on Pentium III 1GHz for each chunk, depending on the implementation.

Results

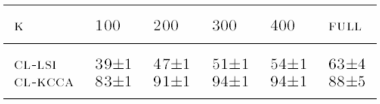
- Pseudo query tests: 5 query words, relevant documents were the test documents themselves in monolinguistic retrieval or their mates in cross-linguistic tests.
- K – unspecified. Probably the number of terms / dimensions.

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



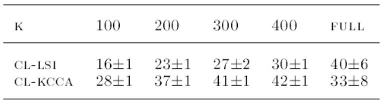
K	100	200	300	400	FULL
CL-LSI	17±1	24±1	28±1	31±1	40±3
CL-KCCA	40±2	55±2	61±1	64±1	60±6

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



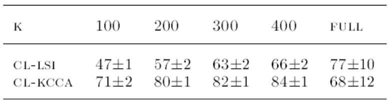
K	100	200	300	400	FULL
CL-LSI	39±1	47±1	51±1	54±1	63±4
CL-KCCA	83±1	91±1	94±1	94±1	88±5

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

Example applications of KCCA

- **Cross-lingual document retrieval**: retrieved documents depend only on the meaning of the query and not its language.
- **Automatic document categorization:** (for example, using SVM.) **Only one** classifier is learned and not a separate classifier for each language.
- **Document clustering**: documents should be grouped into clusters based on their content, not on the language they are written in.
- **Cross-media information retrieval**: in the same way we correlate two languages we can correlate text to images, text to video, text to sound, …

Related approaches

- Usual approach for modelling cross language Information Retrieval is Latent Semantic Indexing (LSI/SVD) on parallel corpora
  - …measured performance of KCCA is consistently and significantly better than LSI on the same data.

[Vinokourov et. al, 2002]

Links

- KCCA is available within Text-Garden text-mining software environment
  - …available at http://www.textmining.net