Learning in Cognitive Systems: Inference of Representations & Grounding through Interaction

John Shawe-Taylor
Information: Signals, Images Systems Group
School of Electronics and Computer Science
University of Southampton

University of Southampton

PASCAL
Pattern Analysis, Statistical Modelling and Computational Learning

Electronics & Computer Science
University of Southampton

Acknowledgements

- Including work (and slides) from
  - Nello Cristianini
  - Jason Farquhar
  - Blaž Fortuna
  - David Hardoon
  - Yaoyong Li
  - Huma Lodhi
  - Anders Meng
  - Craig Saunders
  - Yoram Singer
  - Sandor Szedmak
  - Alexei Vinokourov

Acknowledgements

- Many collaborators and support from EU projects/networks:
  - NeuroCOLT 1 & 2 Networks
  - KerMIT project
  - LAVA project
  - PASCAL Network

Cognitive Systems

- Interact with their environment – with some purpose/function
- Form their own internal representations of the environment – representation grounded in the environment
- Act/output to the environment – in response to their understanding of the environment and of their purpose/function

Machine learning contributions

- Adaptive analysis of input stream data, eg document analysis, image processing, adaptive signal processing
- Integrating/fusing information from different sources, learning compact and therefore semantically informed and potentially useful representations
- Learning to infer from partial/probabilistic observations
- Adaptive control in handling results of its own behaviour
- Statistical/game theoretic analysis of learning and inference

Machine learning contribution

- Adaptive analysis of input stream data, eg document analysis, image processing, adaptive signal processing
- Integrating/fusing information from different sources, learning compact and therefore semantically informed and potentially useful representations
- Learning to infer from partial/probabilistic observations
- Adaptive control in handling results of its own behaviour
- Statistical/game theoretic analysis of learning and inference

Document analysis

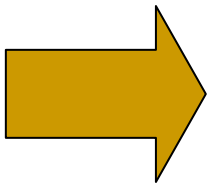
- Provides a good testbed for the ideas – relatively clear ideas of what semantics are and relatively objective measures of performance.
- Standard representation for text documents is the so-called ‘bag-of-words’ mapping:

Bag-of-Words Representation

- Bag of words model – Vector of term weights

$$\mathbf{x} \in \Re^n$$

The higher minimum wage signed into law… will be welcome relief for millions of workers … . The 90-cent-an-hour increase for … .



for	2
into	1
law	1
the	2
.	.
.	.
wage	1

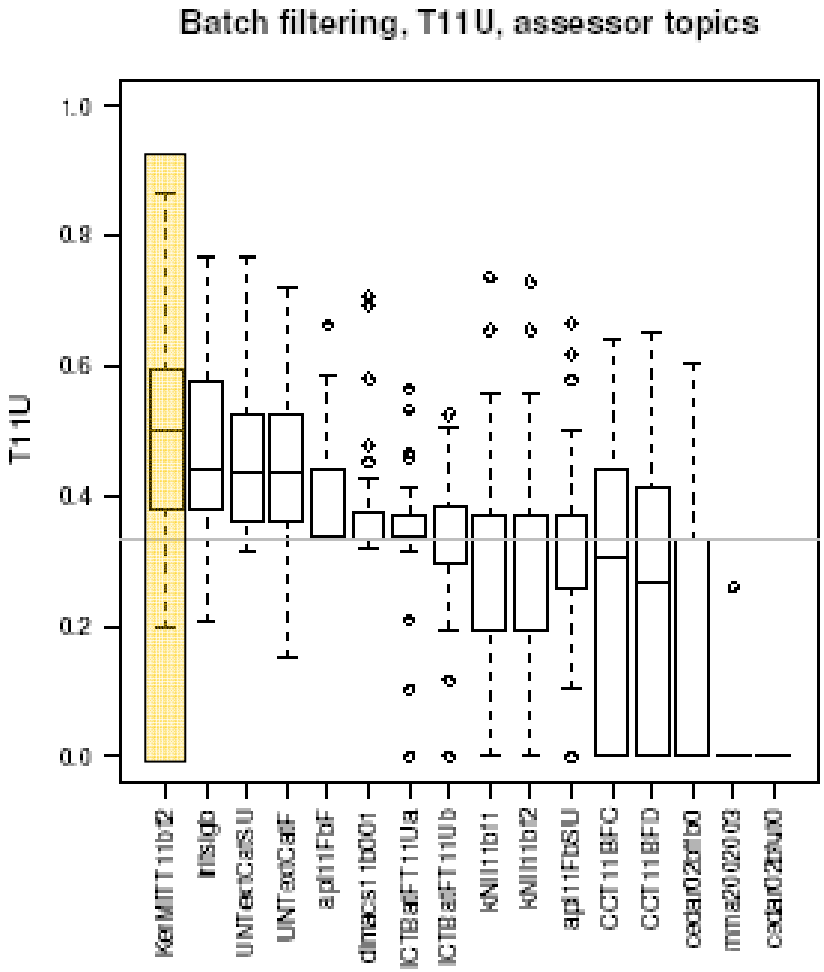
Bag of Words and Kernel Methods

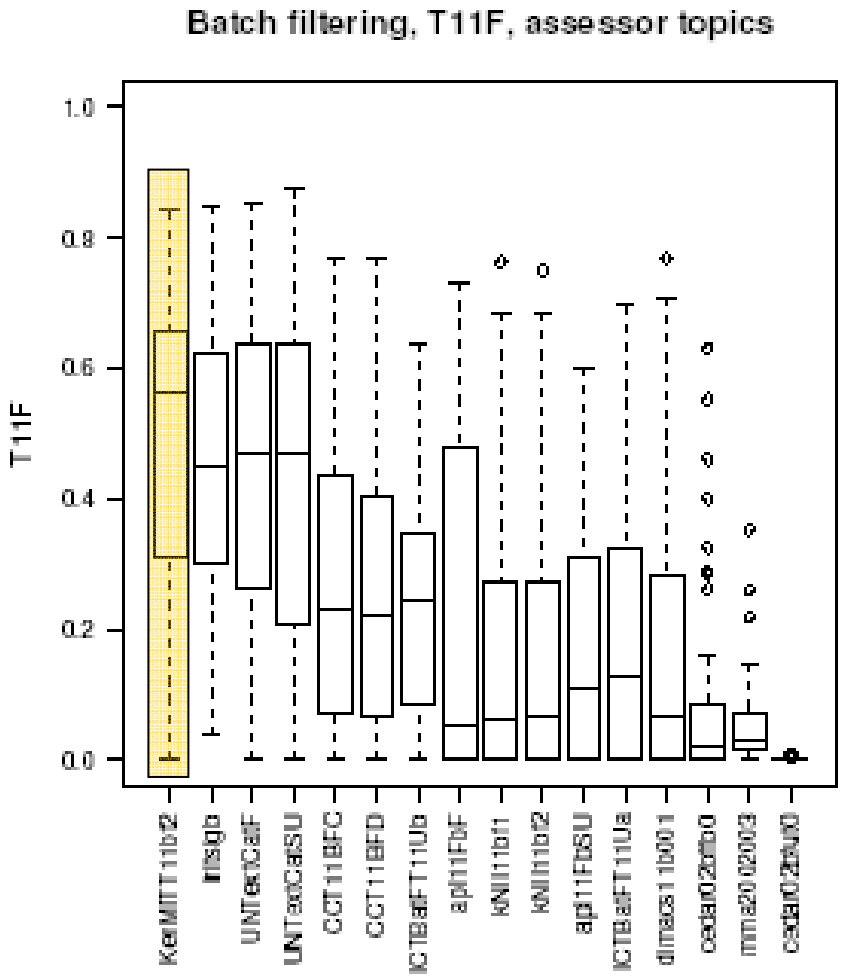
- Kernel methods work with linear functions in high-dimensional feature spaces
- Control flexibility by regularising the norm rather than fixing a (low) dimension
- BoW is a high-dimensional representation – kernel methods work well (eg SVMs) provided have large training sets

Results for document classification

- The baseline for document classification has become the Support Vector Machine
- Many variants: τ-perceptron, second order perceptron, multi-class SVM, ranking, etc.
- These approaches all deliver similarly impressive performance
- Some extensions for hierarchical outputs by learning a maximally discriminative Markov network over tree structure

Results: Batch Filtering





Max margin Markov

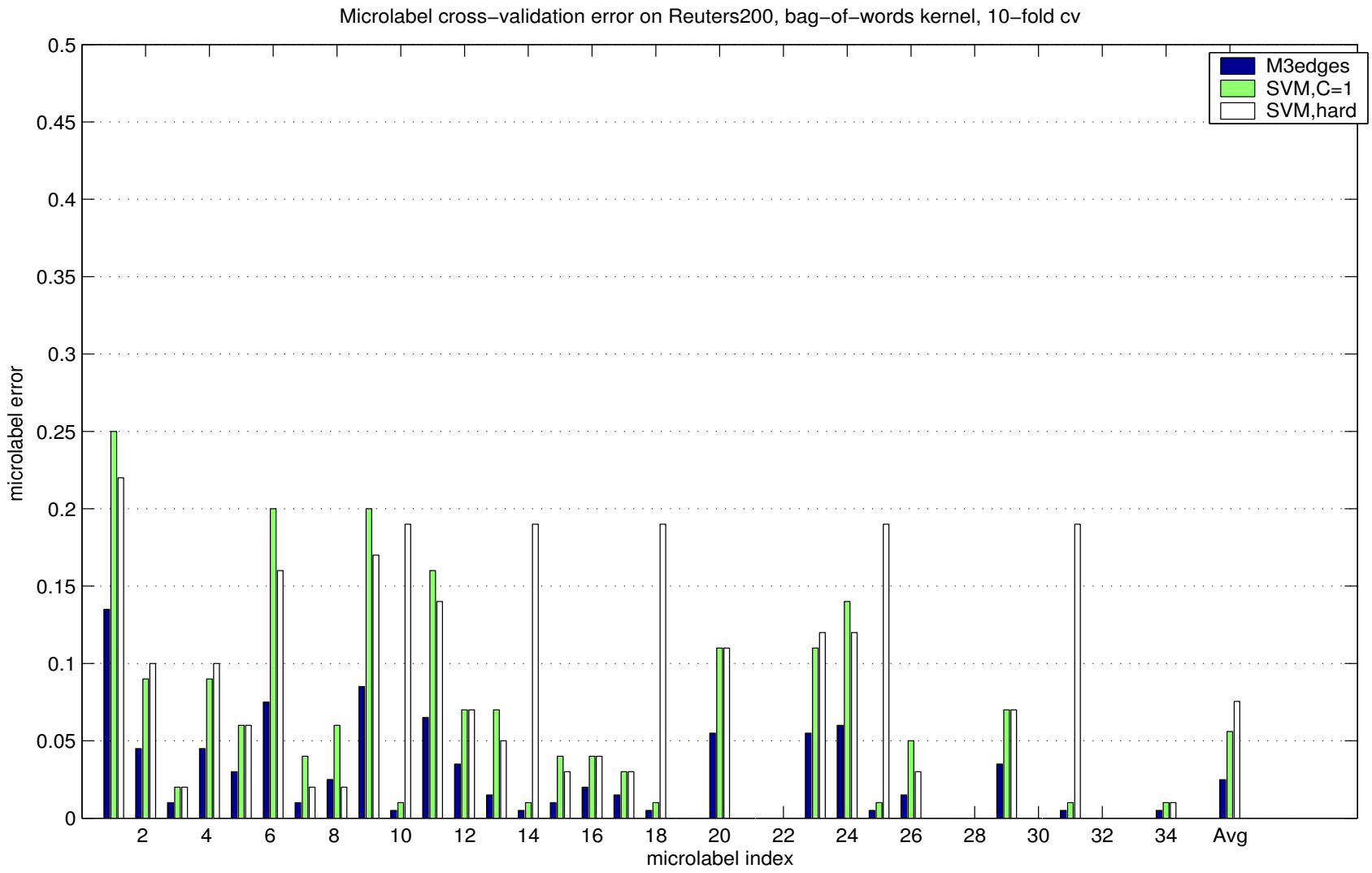


Image processing applications

- Object detection

Representation of images

- Interest points are local features that can be identified in a pre-processing phase
- Some further processing is needed to ‘discretise’ the interest points to create a ‘vocabulary’
- We now view the image as a bag of interest points in the same way as a document is a bag of words
- This has delivered a significant improvement in generic object detection

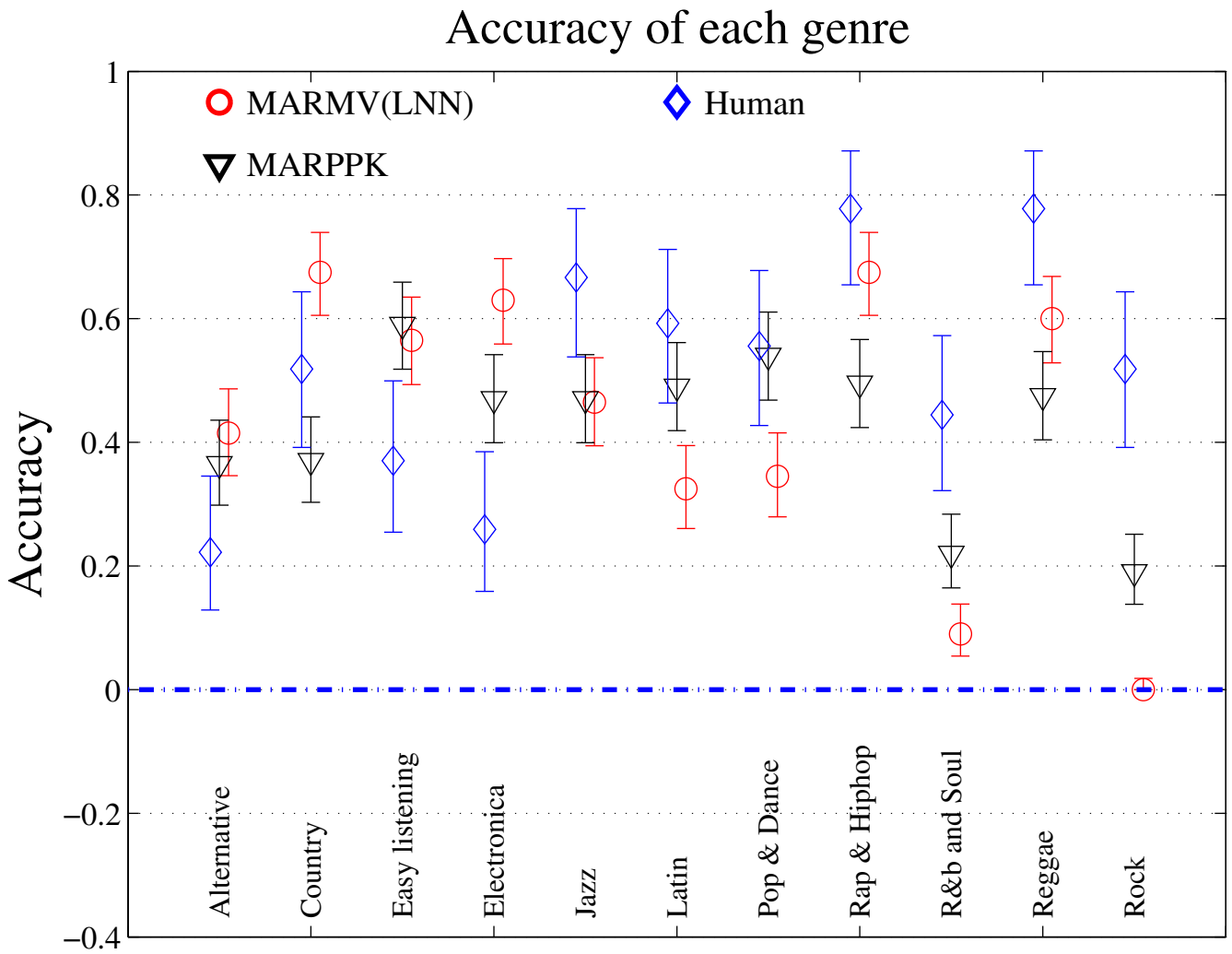
Results for accuracy

	Motorbike	Bicycle	People	Car
SVM	94.05	91.58	91.58	87.95

Signal processing

- Genre classification from audio recordings
- Local (short timescale) auto-regressive modelling gives parameters that can again be bagged to create a representation of a recording
- Apply an SVM or similar learning algorithm
- Results of two combinations MARMV and MARPPK are shown

Results for Genre classification



Machine learning contribution

- Adaptive analysis of input stream data, eg document analysis, image processing, adaptive signal processing
- Integrating/fusing information from different sources, learning compact and therefore semantically informed and potentially useful representations
- Learning to infer from partial/probabilistic observations
- Adaptive control in handling results of its own behaviour
- Statistical/game theoretic analysis of learning and inference

Underlying drivers

- To what extent can one learn semantics from data?
- Do we need handcrafted tools such as wordnet to infer semantics?
- Can we gain insights into how semantics can arise from raw data analysis?
- Can we make similar inroads in semantics as have been made in classification through SLT?
- Will this provide insights into the grounding issue?

BoW matrix

- Let D be the document-term matrix:

$$D = \overbrace{\left.\begin{pmatrix} \mathrm{tf}_{11} & \cdots & \mathrm{tf}_{1j} & \cdots & \mathrm{tf}_{1N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{tf}_{i1} & \cdots & \mathrm{tf}_{ij} & \cdots & \mathrm{tf}_{iN} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathrm{tf}_{m1} & \cdots & \mathrm{tf}_{mj} & \cdots & \mathrm{tf}_{mN} \end{pmatrix}\right\}}^{words} docs$$

where tf_{ij} is the frequency of term j in document i. Note that representation for doc d_i is ith row vector $\phi(d_i)$ with corresponding kernel

$$\kappa(d, d') = \langle \phi(d), \phi(d') \rangle$$

Semantic Matrix

- Semantics via a matrix S that has ijth entry giving relation between term i and term j creates a semantic spread: $D^{\dagger}=DS$
- Can typically consider $S=RP$, where R gives term weightings and P term proximities
- For example idf weightings: $\mathsf{idf}(w) = \ln\left(\frac{m}{\mathsf{df}(w)}\right)$
- Wordnet distances to determine proximities

Latent Semantic Indexing

- Developed in information retrieval to overcome problem of semantic relations between words in BoWs representation
- Use Singular Value Decomposition of Term-doc matrix:

$$D' = U\Sigma V' = \left(\mathrm{u}_1, \ldots, \mathrm{u}_N\right) \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \mathrm{v}'_1 \\ \ldots \\ \mathrm{v}'_m \end{pmatrix}$$

Lower dimensional representation

- If we truncate the matrix U at k columns we obtain the best k dimensional representation in the least squares sense:

$$D' \approx U_k \Sigma_k V_k' = \left(\mathsf{u}_1, \ldots, \mathsf{u}_k\right) \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \begin{pmatrix} \mathsf{v}_1' \\ \ldots \\ \mathsf{v}_k' \end{pmatrix}$$

- In this case we can write the semantic matrix as

$$S = U_k U_k'$$

semantic proximity learned from data.

Latent Semantic Kernels

- Can perform the same transformation in a kernel defined feature space by performing an eigenvalue decomposition of the kernel matrix (this is equivalent to kernel PCA):

$$\phi(d)U_k = \left(\lambda_i^{-1/2} \sum_{j=1}^{m} (\mathbf{v}_i)_j \, \kappa(d_j, d) \right)_{i=1}^{k},$$

where $\lambda_i, \mathbf{v}_i$ are eigenvalue/vectors of kernel matrix.

Theoretical Analysis

- Can relate the quality of approximation on training data to quality on test data (as measured by the eigenvalues) provided number of dimensions k is small compared to number of training examples m.
- This works independently of the dimension - even in infinite dimensional spaces such as that generated by the Gaussian kernel.
- Hence, showing that the semantic directions found will continue to be relevant on new data drawn from the same distribution.

Related techniques

- Number of related techniques:
  - Probabilistic LSI (pLSI)
  - Non-negative Matrix Factorisation (NMF)
  - Multinomial PCA (mPCA)
  - Discrete PCA (DPCA)
- All can be viewed as alternative decompositions:

$$D' \approx CM \text{ (in LSI/PCA} \approx U_k(\Sigma_k V_k')\text{)}$$

where k columns of C are underlying components and M gives mixing to create different documents.

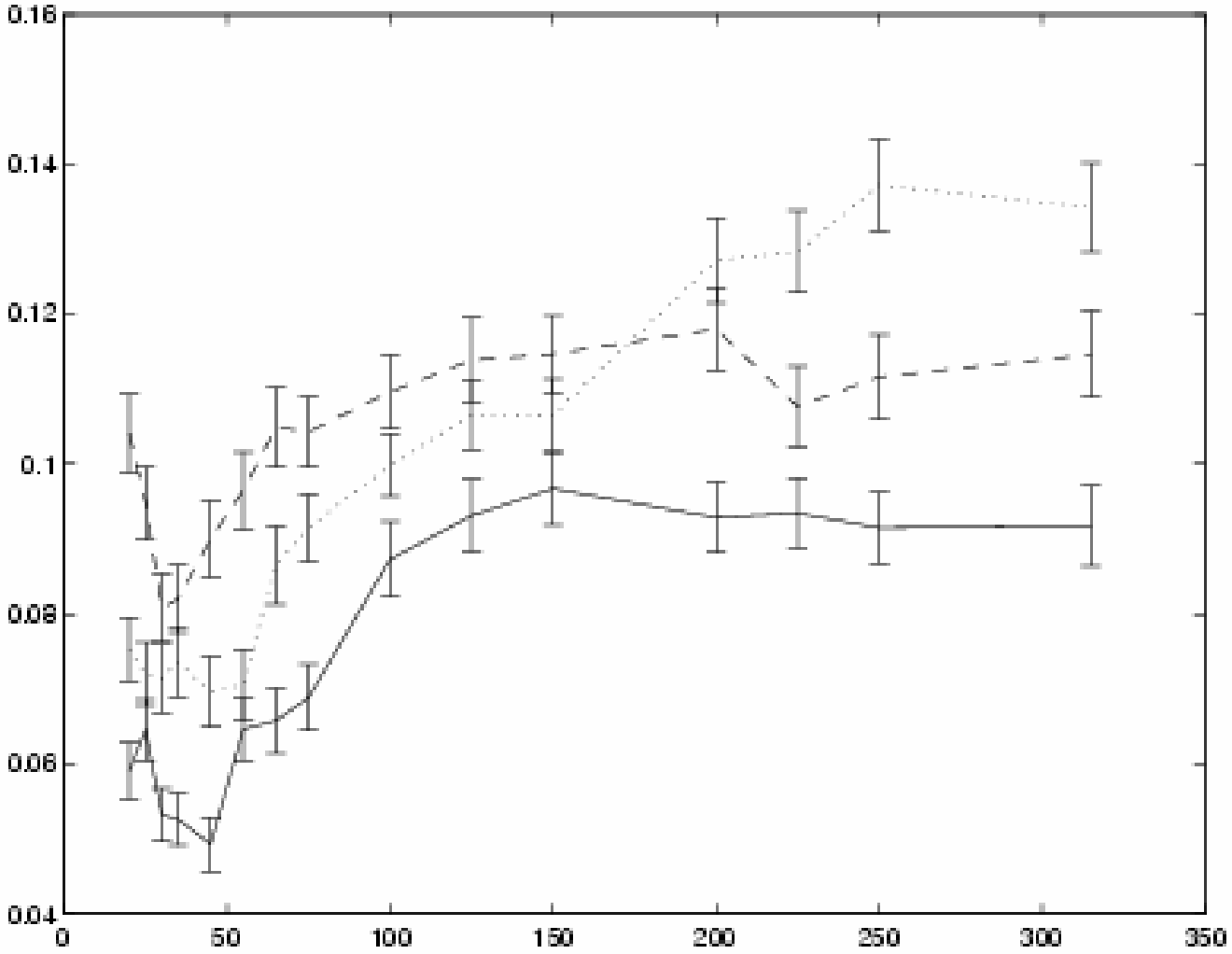
Different criteria

- Vary by:
  - Different constraints (eg non-negative entries)
  - Different prior distributions (eg Dirichlet, Poisson)
  - Different optimisation criteria (eg max likelihood, Bayesian)
- Unlike LSI typically suffer from local minima and so require EM type iterative algorithms to converge to solutions

Using a semantic space

- LSI introduced for Information Retrieval: project query into the semantic space and look for documents that have highest inner product with query in that space
- Can use representation for SVM training – improves performance for small training sets

Generalisation error for Ionosphere data with poly kernels of degrees 2 (unbroken) 3 (dashed) and 4 (dotted)

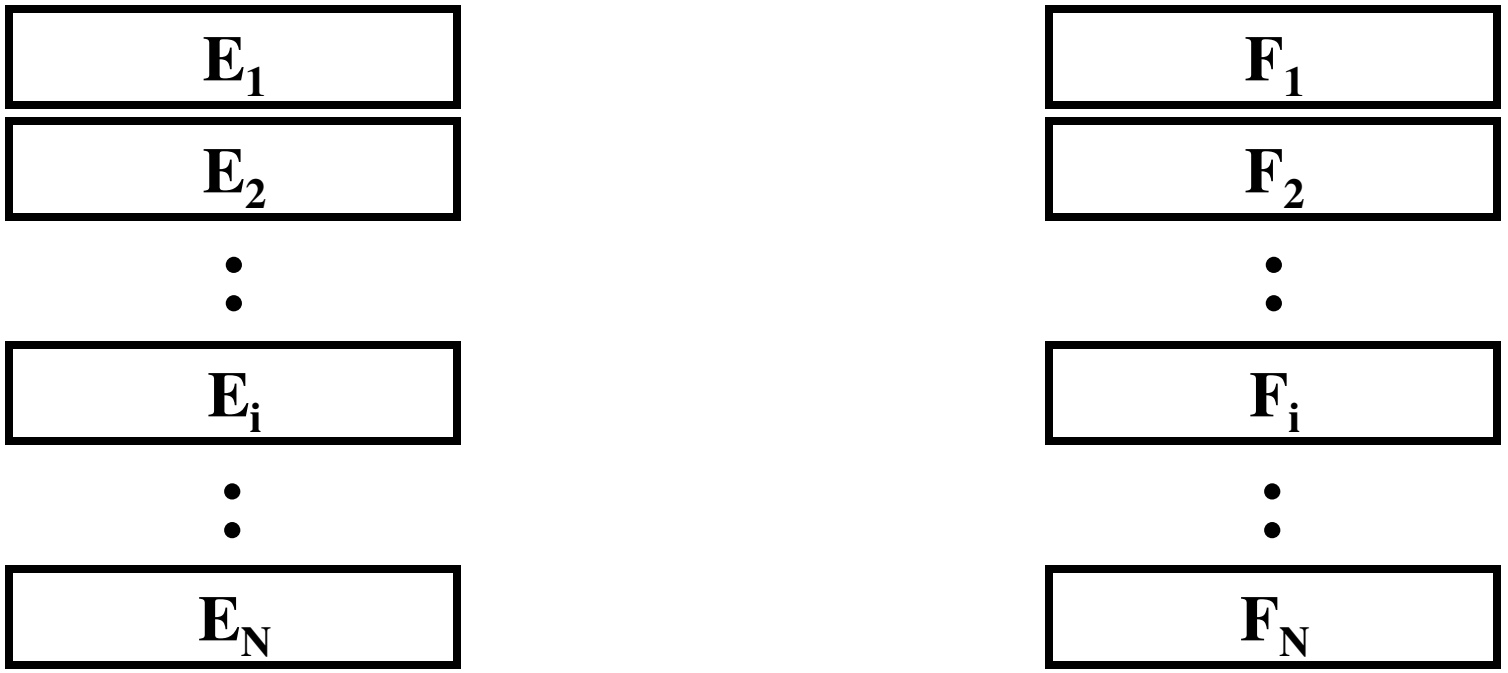


Paired corpora

- Can we use information from paired corpora to extract more information?
- Two views of same semantic object – hypothesise that both views contain all of the necessary information, eg document and translation to a second language:

$$\phi_a(d) \longleftarrow d \longrightarrow \phi_b(d)$$

aligned text



Canadian parliament corpus

LAND MINES
Ms. Beth Phinney (Hamilton Mountain, Lib.):
Mr. Speaker, we are pleased that the Nobel peace prize has been given to those working to ban land mines worldwide.
We hope this award will encourage the United States to join the over 100 countries planning to come to …

E_{12}

LES MINES ANTIPERSONNEL
Mme Beth Phinney (Hamilton Mountain, Lib.):
Monsieur le Président, nous nous réjouissons du fait que le prix Nobel ait été attribué à ceux qui oeuvrent en faveur de l'interdiction des mines antipersonnel dans le monde entier.
Nous espérons que cela incitera les Américains à se joindre aux représentants de plus de 100 pays qui ont l'intention de venir à …

F_{12}

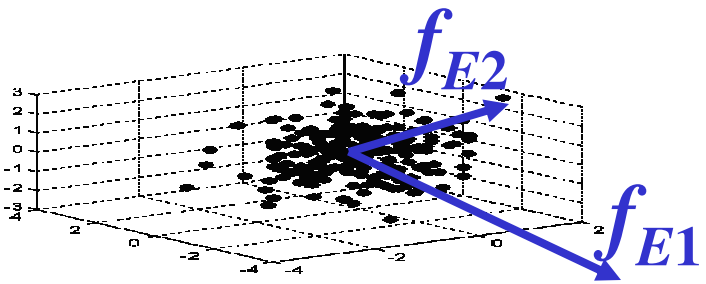
cross-lingual lsi via svd

$$D' = \begin{pmatrix} D'_E \\ D'_F \end{pmatrix} = \begin{pmatrix} U_E \\ U_F \end{pmatrix} \Sigma V^T$$

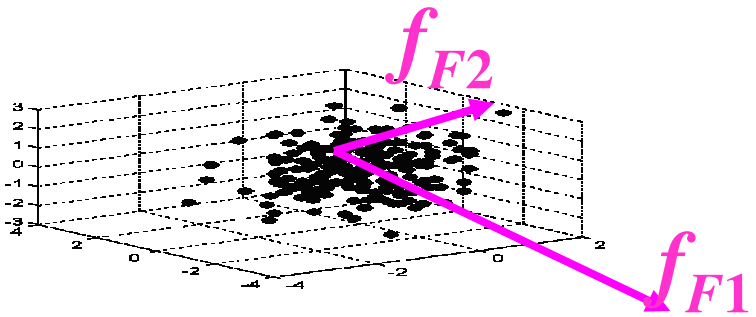
M. L. Littman, S. T. Dumais, and T. K. Landauer. Automatic cross-language information retrieval using latent semantic indexing. In G. Grefenstette, editor, *Cross-language information retrieval*. Kluwer, 1998.

cross-lingual kernel canonical correlation analysis

feature “English” space

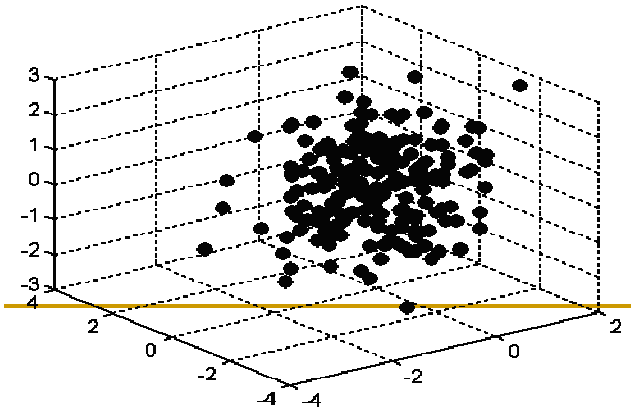


feature “French” space

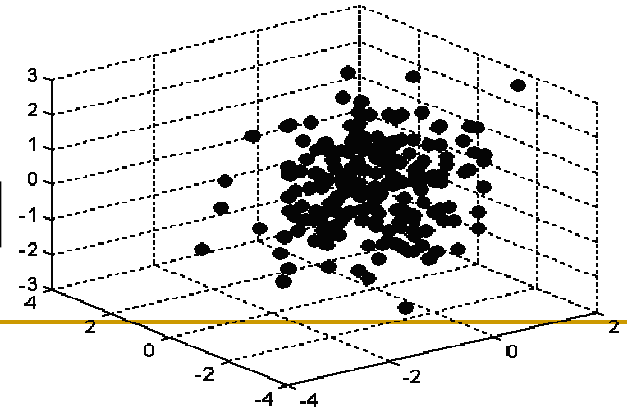


$\Phi(x)$

input “English” space

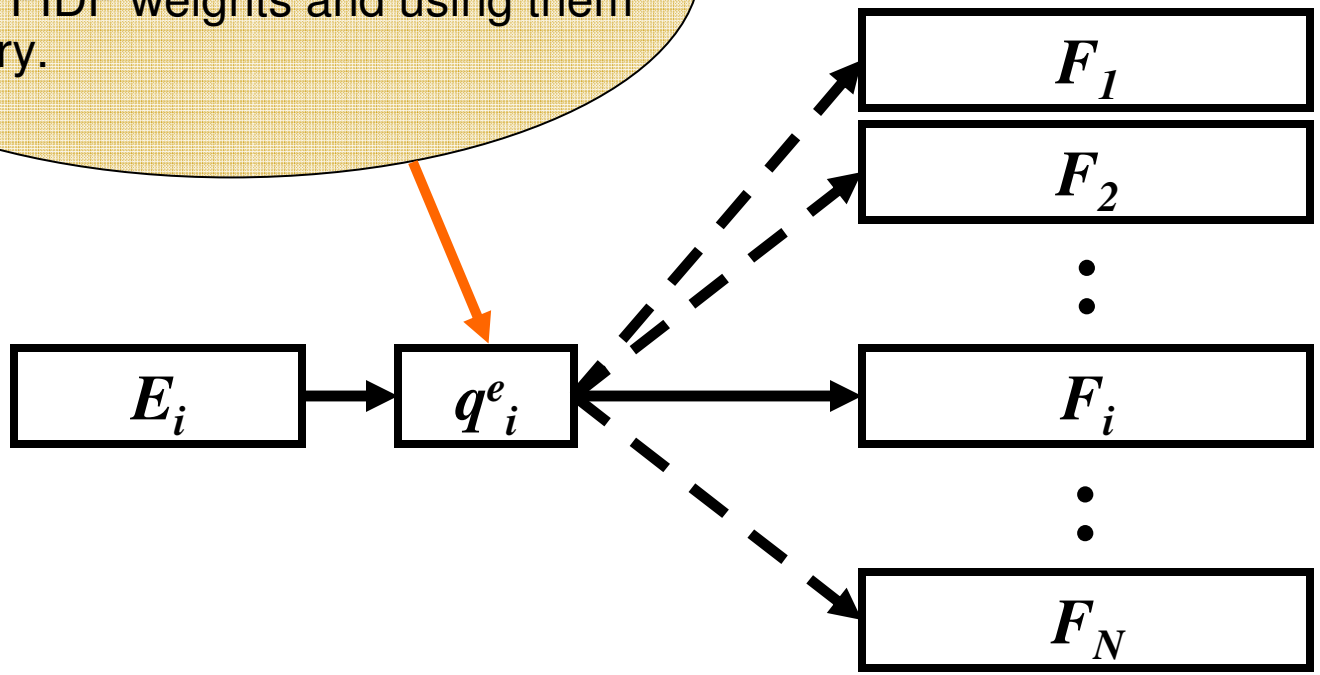


input “French” space



pseudo query test

Queries were generated from each test document by extracting 5 words with the highest TFIDF weights and using them as a query.



Experimental Results

The goal was to retrieve the paired document.

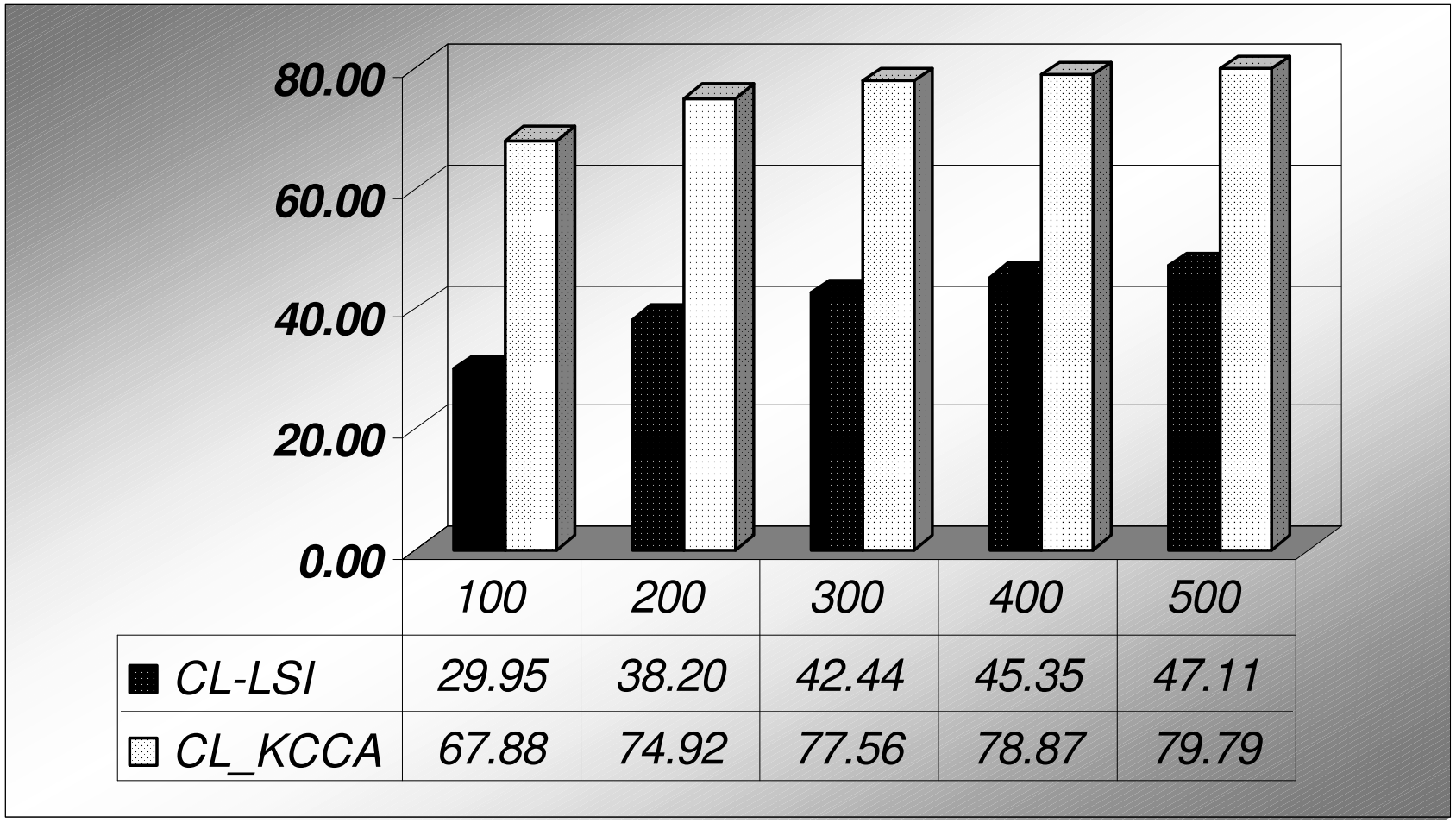
Experimental procedure:

(1) LSI/KCCA trained on paired documents,

(2) All test documents projected into the LSI/KCCA semantic space,

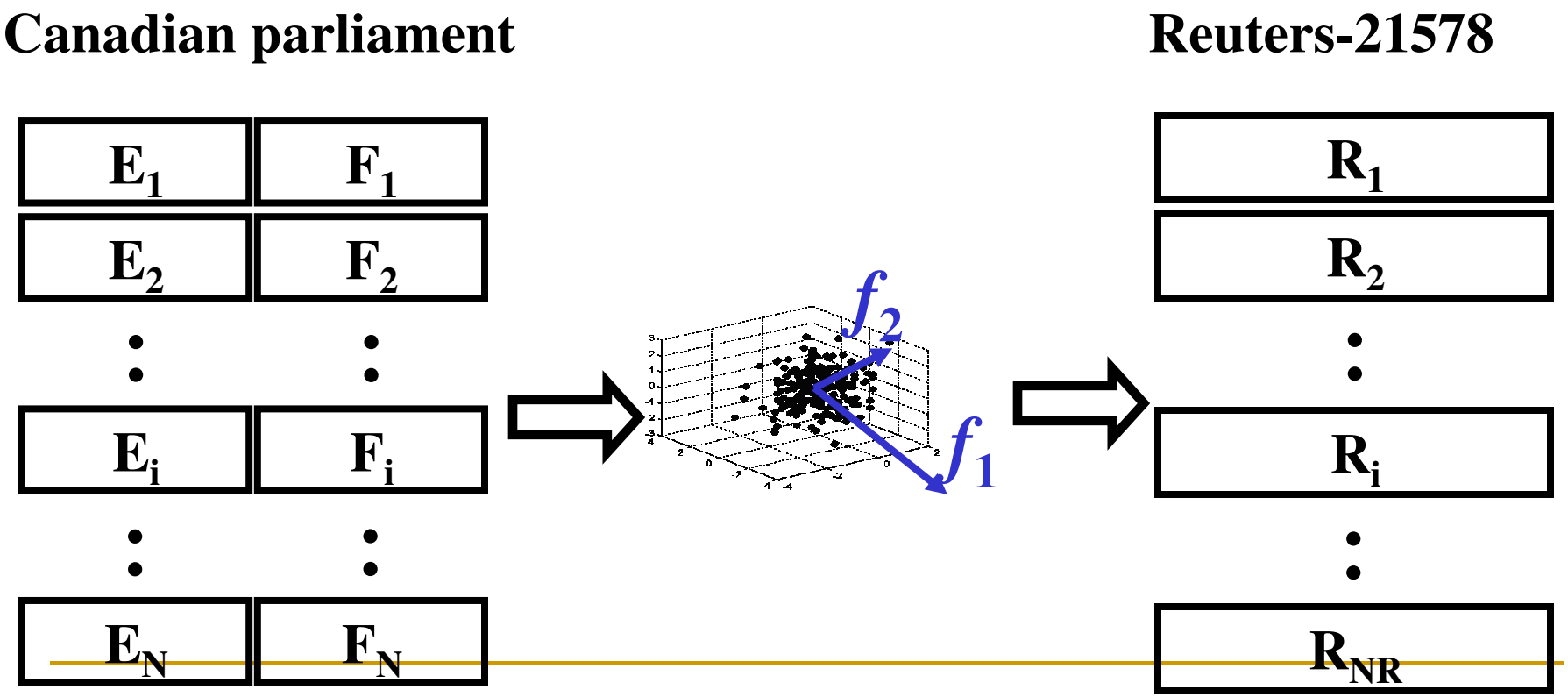
(3) Each query was projected into the LSI/KCCA semantic space and documents were retrieved using nearest neighbour based on cosine distance to the query.

English-French retrieval accuracy, %



	100	200	300	400	500
CL-LSI	29.95	38.20	42.44	45.35	47.11
CL_KCCA	67.88	74.92	77.56	78.87	79.79

using semantics, extracted from aligned corpus, for completely different corpora



Classification Procedure

Experimental procedure:

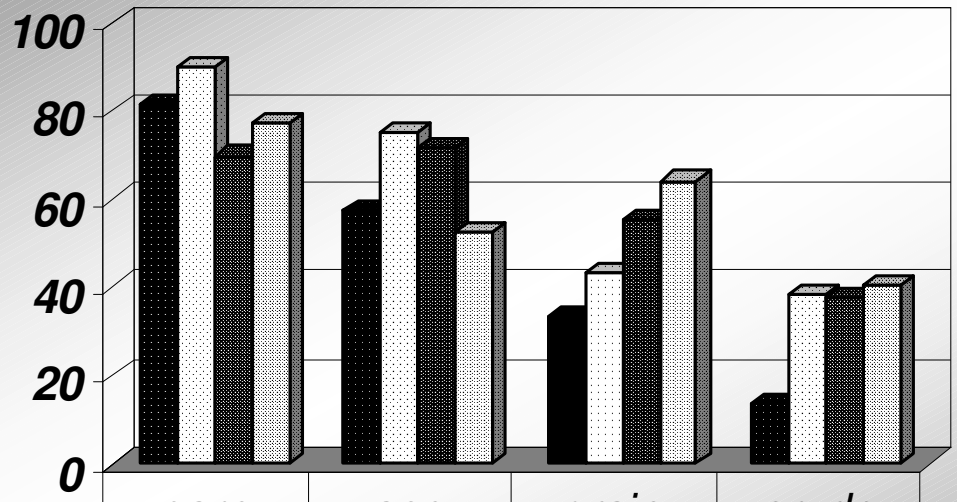
(1) LSI/KCCA trained on paired documents,

(2) Whole Reuters corpus was projected into the LSI/KCCA semantic space,

(3) Linear SVM classifier was trained in the LSI/KCCA semantic space on a subset of documents and tested on a separate test set.

(4) Same procedure used for Generalised Vector Space Model (GVSM) – representation of a document is vector of inner products with the training set of appropriate language.

SVM classification with Reuters 21578 with 5% training

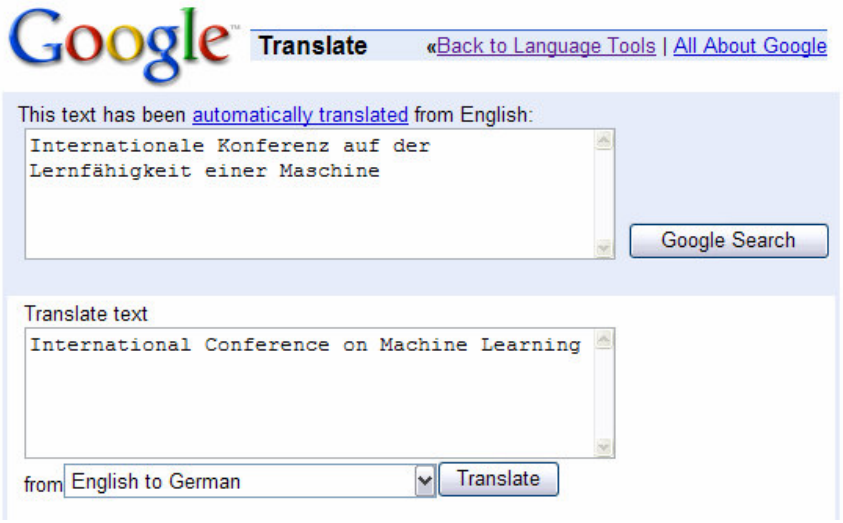


	earn	acq	grain	crude
BAG-OF-WORDS	82	57	34	13
CL-KCCA	90	75	43	38
GVSM	70	72	56	37
CL-LSI	77	52	64	40

Paired training set and machine translation

KCCA needs paired dataset for training. When there is no paired dataset available we have two options:

- We use human made dataset from some other domain.
  - This could be unreliable because of a big semantic and vocabulary gap.
- We use machine translation tools to generate paired dataset.
  - In our experiments we used Google Language Tools for translating documents.



Artificial paired corpora

We compared two paired corpora:

- *Hansard* corpus: aligned pairs of text chunks from the official records of the 36th Canadian Parliament Proceedings.

  [Germann, 2001]

- *Artificial corpus*: half of the English and half of the French translations from *Hansard* corpus were replaced by machine translation.

Results

For 65% of queries the correct document appeared on the first place.
For 95% of queries the correct document appeared among first 10 results.

	En-En	En-Fr	Fr-En	Fr-Fr
Hansard	87 / 99	66 / 96	65 / 95	84 / 99
Artificial	86 / 99	58 / 91	59 / 90	83 / 99

There is no difference when query and document are in the same language

When query and document are from different languages, there is around 5-10% drop in retrieval accuracy

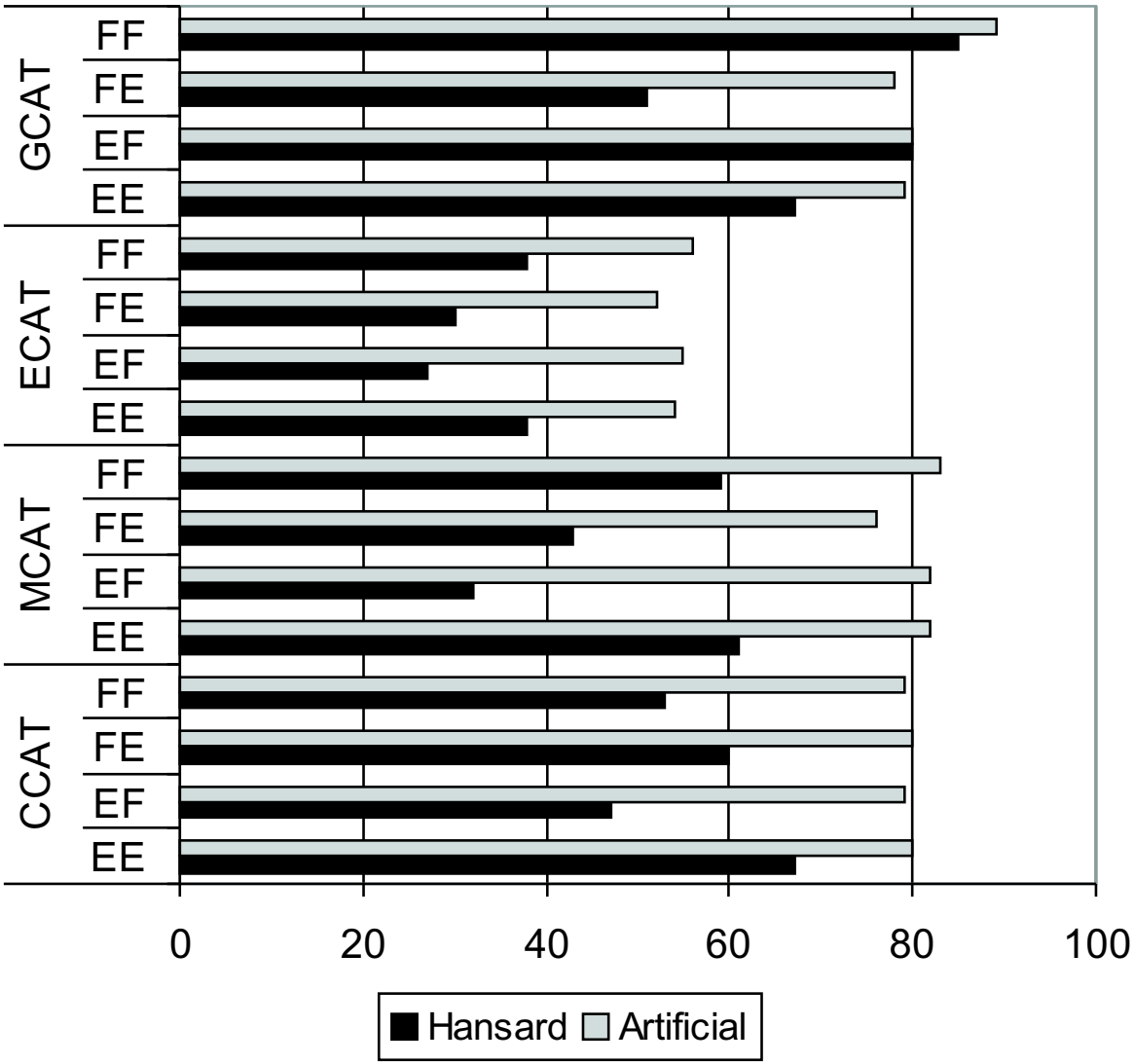
Classification with artificial PC

Reuters multilingual corpora (English and French) was used as a dataset. [Reuters, 2004]

- First paired train set, *Hansard*, was taken from previous experiment; different domain than news articles.
- Second paired train set was generated from the Reuters dataset using machine translation (Google).

- Results are averaged over 5 random splits.

Results



#KCCA dimensions: 800

FE … French training set, English testing set.

Artificial paired training set generates significantly better semantic space than train set taken from a different domain!

Applying to different data types

- Data
  - Combined image and associated text obtained from the web
  - Three categories: sport, aviation and paintball
  - 400 examples from each category (1200 overall)
  - Features extracted: HSV, Texture , Bag of words
- Tasks
  1. Classification of web pages into the 3 categories
  2. Text query -> image retrieval

Classification error of baseline method

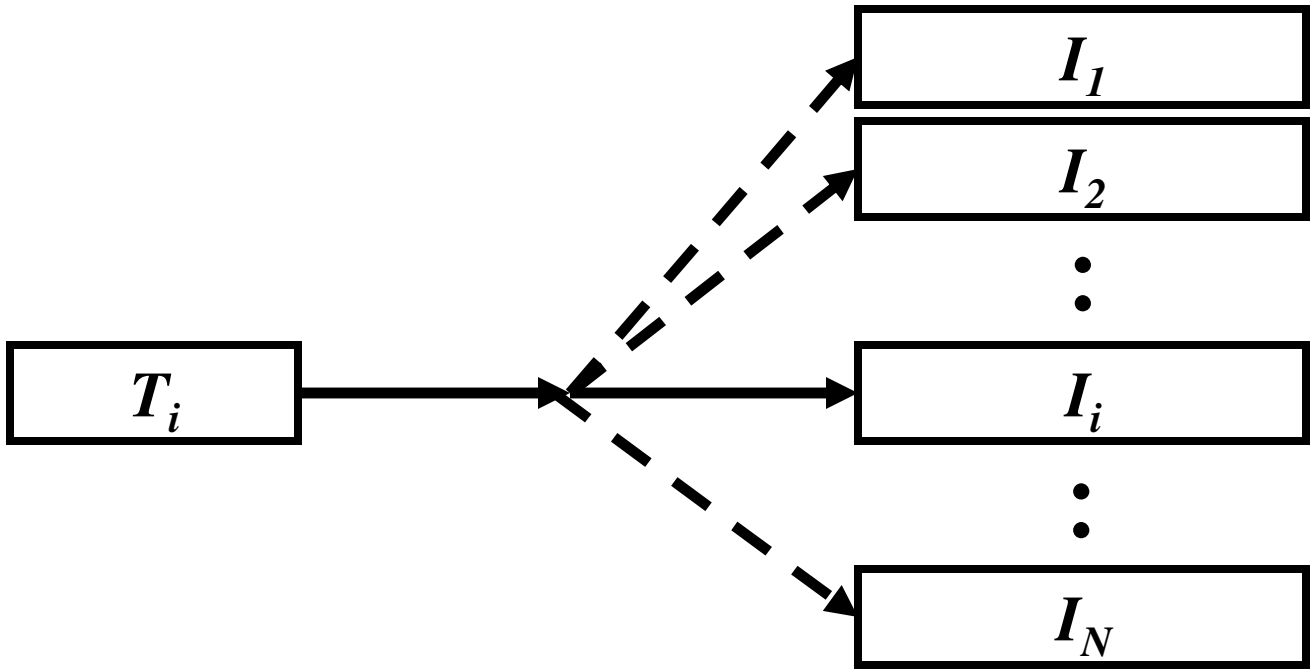
- Previous error rates obtained using probabilistic ICA classification done for single feature groups

Colour	Texture	Colour + Texture	Text	All
22.9%	18.3%	13.6%	9.0%	3.0%

Classification rates using KCCA

K	Error rate
Plain SVM	2.13%±0.23%
KCCA-SVM (150)	1.36%±0.15%
KCCA-SVM (200)	1.21%±0.27%

Query Test for Image Retrieval



% Success of partner image in Top n rated images

Success rate	Top 10	Top 30
GVSM	1.27%	5%
KCCA	17.34%	30.34%

Example

Height: 7-0 weight: 225 lbs position: center born: august 5, 1962, Kingston, Jamaica college: Georgetown







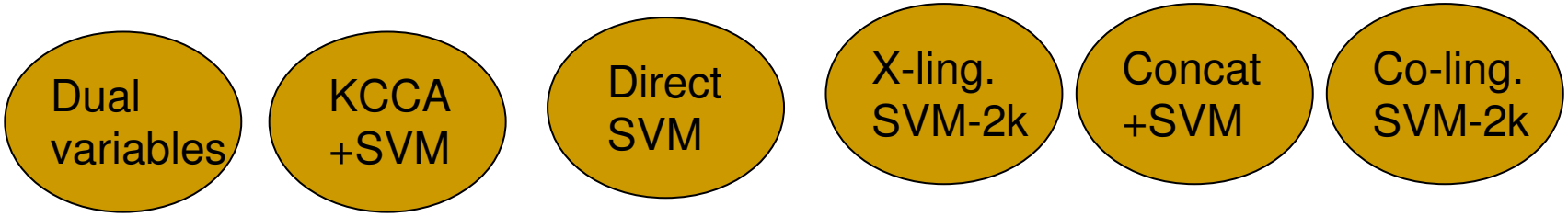



Actual Match

Classification with multi-views

- If we use KCCA to generate a semantic feature space and then learn with an SVM, can envisage combining the two steps into a single ‘SVM-2k’
- learns two SVMs one on each representation, but constrains their outputs to be similar across the training (and any unlabelled data).
- Can give classification for single mode test data – applied to patent classification for Japanese patents
- Again theoretical analysis predicts good generalisation if the training SVMs have a good match, while the two representations have a small overlap

Results



	pSVM	kcca_SVM	SVM	SVM_2k_j	Concat	SVM_2k
1	59.4±3.9	60.3±2.8	66.6±2.8	66.1± 2.6	67.5±2.3	67.5±2.1
2	71.1±4.5	68.4±4.4	73.0±4.0	74.8±4.7	73.9±4.0	75.1±4.1
3	16.7±1.2	13.1±1.0	18.8±1.6	20.8±1.9	21.5±1.9	22.5±1.7
7	74.9±1.8	76.0±1.2	76.7±1.3	77.5±1.4	79.0±1.2	80.7±1.5
12	75.0±0.8	73.6±0.8	76.8±1.0	77.6±0.7	76.8±0.6	78.4±0.6
14	76.0±1.6	71.5±1.5	80.9±1.3	82.2±1.3	81.4±1.4	82.7±1.3

Learned representations are grounded?

- For example we can view Pavlov's experiments with conditioning as training the dog to create a new representation for eating by correlating the delivery of food with the opening of the door
- The SVM-2K model suggests that this can trigger anticipation of eating when just one representation is triggered.

Grounded representations and experience

- Learning from data means that any symbol processing is secondary – the primary link between the source or sources is through internally developed representations
- Learning through correlations leads us to more compact and hence semantically informed representations that render the task of acting/manipulating the environment simpler
- Suggests that there is a need to understand how relevant semantics/low dimensional representations can be inferred that retain the right level of flexibility
- There is a need for a theoretical framework to analyse and guide the design of these systems and associated algorithms

Emergent properties

- Many of the components mentioned above have been developed as part of past/current EU and other projects
- Challenge/opportunity to use these components to build a system from many small adaptive agents, creating a complex system with emergent properties
- Ideal domain for genuine cognitive component – web: computer readable information with diverse sources, content, new scale of data processing, etc.
- Cognition will be an emergent property of the system as it delivers suitably integrated content adapted to different users' interests, etc.

How far can data-driven approaches go?

- Recent developments of statistical machine translation suggests that even the most complex tasks are amenable to the approach
- The approach can form the basis for the development of complex cognitive systems with many distributed components performing analysis tasks independently
- Results stored in inferred representations grounded in the ‘experience’ of the system
- Adaptation of content for individual users based on inferred preferences is just one more on-line learning module

Impact

- Cognitive systems hold the potential to deliver a new level of intelligent processing in virtually all walks of life
- They can meet the challenge of information overload on the internet and work
- Intelligent behaviour in transport networks, etc.
- Handling complexity in a range of contexts, eg air traffic control, environment, etc.

Conclusions

- Machine Learning has already begun to play an important role in improving processing of a wide variety of data
- We have argued that it can play an equally important role in inferring grounded semantically relevant representations
- The time would appear ripe for an attempt to create a complex cognitive system following this paradigm, perhaps focussed on web analysis
- There is a need for a better theoretical framework within which to analyse and hence inform the design of such systems