

# Correlation for multilingual analysis: Theory, Scaling and Sparsity

John Shawe-Taylor

Department of Computer Science  
University College London

MLIA Workshop, NIPS 2006

Joint work with David Hardoon, Zhuoran Wang and Sandor  
Szedmak

- Canonical Correlation Multi-lingual Analysis
- Cross-lingual information retrieval and classification
- SVM-2K: theory and practice
- Scaling KCCA to large datasets: primal and dual sparsity
- Regression Framework for Statistical Machine Translation
- Discussion & Future Work

# Inferring semantic models from two views

- Consider two different views of the same object:

$$\mathbf{x} \mapsto (\phi_a(\mathbf{x}), \phi_b(\mathbf{x}))$$

- Examples:
  - documents in two languages
  - image and its caption
  - brain scan and corresponding activity description
- Would like to use the 'triangulation' provided by the two views to infer a refined model of the underlying semantics

# Canonical correlation analysis

- Assume that the semantics can be expressed as a linear combination of the views
- CCA seeks features that are maximally correlated
- Seek  $\mathbf{w}_a$  creating feature  $x_a = \mathbf{w}'_a \phi_a(\mathbf{x})$  and  $\mathbf{w}_b$  creating feature  $x_b = \mathbf{w}'_b \phi_b(\mathbf{x})$  that maximise:

$$\rho = \frac{\hat{\mathbb{E}}[x_a x_b]}{\sqrt{\hat{\mathbb{E}}[x_a^2] \hat{\mathbb{E}}[x_b^2]}} = \frac{\hat{\mathbb{E}}[\mathbf{w}'_a \phi_a(\mathbf{x}) \phi_b(\mathbf{x})' \mathbf{w}_b]}{\sqrt{\hat{\mathbb{E}}[(\mathbf{w}'_a \phi_a(\mathbf{x}))^2] \hat{\mathbb{E}}[(\mathbf{w}'_b \phi_b(\mathbf{x}))^2]}}$$

- Consider the notation  $\mathbf{X}_a, \mathbf{X}_b$  for the matrices with feature vectors for the two views of an object in corresponding rows.

- Using this notation

$$\rho = \frac{\mathbf{w}'_a \mathbf{X}'_a \mathbf{X}_b \mathbf{w}_b}{\sqrt{\mathbf{w}'_a \mathbf{X}'_a \mathbf{X}_a \mathbf{w}_a \mathbf{w}'_b \mathbf{X}'_b \mathbf{X}_b \mathbf{w}_b}}$$

- Since invariant to rescalings we can set two factors in denominator equal to 1. Using Lagrange multipliers obtain  $L(\mathbf{w}_a, \mathbf{w}_b, \lambda_a, \lambda_b)$  as

$$\mathbf{w}'_a \mathbf{X}'_a \mathbf{X}_b \mathbf{w}_b - \frac{\lambda_a}{2} \mathbf{w}'_a \mathbf{X}'_a \mathbf{X}_a \mathbf{w}_a - \frac{\lambda_b}{2} \mathbf{w}'_b \mathbf{X}'_b \mathbf{X}_b \mathbf{w}_b$$

- Gives coupled equations

$$\mathbf{X}'_a \mathbf{X}_b \mathbf{w}_b = \lambda_a \mathbf{X}'_a \mathbf{X}_a \mathbf{w}_a \text{ and } \mathbf{w}'_a \mathbf{X}'_a \mathbf{X}_b = \lambda_b \mathbf{w}'_b \mathbf{X}'_b \mathbf{X}_b$$

- Simple to verify that  $\lambda_a = \lambda_b$
- Resulting in generalised eigenvalue problem:

$$\begin{pmatrix} \mathbf{0} & \mathbf{X}'_a \mathbf{X}_b \\ \mathbf{X}'_b \mathbf{X}_a & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{X}'_a \mathbf{X}_a & \mathbf{0} \\ \mathbf{0} & \mathbf{X}'_b \mathbf{X}_b \end{pmatrix} \begin{pmatrix} \mathbf{w}_a \\ \mathbf{w}_b \end{pmatrix}$$

# Canonical correlation analysis

- Danger of overfitting since in high-dimensional spaces we can find a perfect fit for  $\mathbf{w}_b$  whatever  $\mathbf{w}_a$  we choose.
- Consider the function:

$$g_{\mathbf{w}_a, \mathbf{w}_b}(\mathbf{x}) = \left\| \mathbf{w}_a^T \phi_a(\mathbf{x}) - \mathbf{w}_b^T \phi_b(\mathbf{x}) \right\|^2,$$

if  $\|\mathbf{w}_a\| \leq A$  and  $\|\mathbf{w}_b\| \leq B$  we can bound

$$\begin{aligned} \mathbb{E}[g_{\mathbf{w}_a, \mathbf{w}_b}(\mathbf{x})] &\leq \sum_{i=1}^k 2(1 - \rho_i) + 3R(A^2 + B^2) \sqrt{\frac{\ln(\frac{2}{\delta})}{2\ell}} \\ &\quad + \frac{4(A^2 + B^2)k}{\ell} \sqrt{\sum_{i=1}^{\ell} (\kappa_a(x_i, x_i) + \kappa_b(x_i, x_i))^2} \end{aligned}$$

- We would like to make a kernel version of this procedure – but must ensure that the flexibility is controlled to rule out spurious correlations
- As shown above another application of the regularisation strategy:

$$\rho = \max_{\mathbf{w}_a, \mathbf{w}_b} \mathbf{w}'_a \mathbf{X}'_a \mathbf{X}_b \mathbf{w}_b$$

subject to:  $(1 - \tau) \mathbf{w}'_a \mathbf{X}'_a \mathbf{X}_a \mathbf{w}_a + \tau \mathbf{w}'_a \mathbf{w}_a = 1$

and  $(1 - \tau) \mathbf{w}'_b \mathbf{X}'_b \mathbf{X}_b \mathbf{w}_b + \tau \mathbf{w}'_b \mathbf{w}_b = 1$

- We now dualise by letting  $\mathbf{w}_a = \mathbf{X}'_a \alpha$  and  $\mathbf{w}_b = \mathbf{X}'_b \beta$  to obtain:

$$\begin{aligned} \rho &= \max_{\alpha, \beta} \alpha' \mathbf{K}_a \mathbf{K}_b \beta \\ \text{subject to:} & \quad (1 - \tau) \alpha' \mathbf{K}_a^2 \alpha + \tau \alpha' \mathbf{K}_a \alpha = 1 \\ \text{and} & \quad (1 - \tau) \beta' \mathbf{K}_b^2 \beta + \tau \beta' \mathbf{K}_b \beta = 1 \end{aligned}$$

- giving the generalised eigenvalue problem

$$\begin{aligned} & \begin{pmatrix} \mathbf{0} & \mathbf{K}_a \mathbf{K}_b \\ \mathbf{K}_b \mathbf{K}_a & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \\ &= \lambda \begin{pmatrix} (1 - \tau) \mathbf{K}_a^2 + \tau \mathbf{K}_a & \mathbf{0} \\ \mathbf{0} & (1 - \tau) \mathbf{K}_b^2 + \tau \mathbf{K}_b \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \end{aligned}$$

# Space identified by CCA

- Subspace identified by CCA should capture the common semantics of the two views – effectively cleaning each of any view-specific noise.
- The subspace projection can then be used for different types of further processing.
- An example with English/Japanese documents performing mate retrieval (1000 training documents): the accuracy rates averaged over 2000 test documents.

#Eigenvectors	5	10	50	100	200	500	1000
Test docs as queries							
KCCA(E→J )	0.050	0.154	0.401	0.471	0.534	0.486	0.377
KCCA(J→E)	0.084	0.173	0.369	0.448	0.461	0.369	0.272
LSI(E→J )	0.037	0.095	0.296	0.376	0.431	0.393	0.247
LSI(J→E)	0.029	0.079	0.212	0.294	0.362	0.304	0.170

# SVM-2k: definition

- In the patent example we could envisage a two-phase process: first identify the subspace using KCCA and then train an SVM on the elicited features to perform cross-lingual classification.
- SVM-2k attempts to combine these two stages into one.
- Train two linear functions (SVMs)  $f_a$  and  $f_b$  on the two views, but add a further series of constraints:

$$|f_a(x_i) - f_b(x_i)| \leq \epsilon + \eta_i, i = 1, \dots, m$$

and add  $D \sum_{i=1}^m \eta_i$  to the objective.

Note: this is before thresholding.

- If observe that  $\sum_{i=1}^m \eta_i$  is small, can estimate the Rademacher complexity of pairs of functions with similarly constrained correlations.
- This leads to an optimisation problem that must be solved to estimate the value of the complexity.
- In practice significant reductions in complexity are achieved with corresponding improvement in classification accuracy.

# SVM-2k: results

- Results obtained classifying patents from a Japanese patent dataset with paired English translations.
- Results are average precision as percent – i.e. higher is better.
- Note that SVM-2k<sub>j</sub> is performing crosslingual classification, i.e. only uses Japanese text – and often does better than an SVM in the original language

	pSVM	kcca_SVM	SVM	SVM-2k <sub>j</sub>	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

# Primal-dual sparsity

$$\begin{aligned}\max_{\mathbf{w}, \mathbf{e}} \quad & -\tau^2 \mathbf{w}' X X' \mathbf{w} + 2\tau(1 - \tau) \mathbf{w}' X K_b \mathbf{e} - (1 - \tau)^2 \mathbf{e}' K_b^2 \mathbf{e} \\ & = -\|\tau X' \mathbf{w} - (1 - \tau) K_b \mathbf{e}\|_2^2\end{aligned}$$

subject to

$$\|\mathbf{w}\|_1 \leq C, \quad \|\mathbf{e}\|_1 = C, \quad \mathbf{e}_i \geq 0$$

and  $0 \leq \tau \leq 1$  can be chosen to ensure that  $\mathbf{w}' X X' \mathbf{w} = \mathbf{e}' K_b^2 \mathbf{e}$ .

Rescaling shows that this combination is independent of  $C$ . Note last equality required to force non-trivial solution.

- Optimisation is convex because can be written as minimising the squared norm of a linear function over a convex region.
- Note also that if we rescale  $\mathbf{w}$  obtained for one value of  $\tau$  we will obtain the optimal solution for a different value of  $\tau$ . Hence, we get equivalent solutions by setting  $\tau = 0.5$  and varying the ratio of the 1-norm bounds.
- Don't actually need to have a paired dataset, but for a paired dataset can also implement primal-primal sparsity.

# Primal-dual sparsity

Let us define the following

$$\mathbf{w}^+ - \mathbf{w}^- = \mathbf{w}, \quad \mathbf{w}^+ \geq 0, \quad \mathbf{w}^- \geq 0.$$

The corresponding Lagrangian is

$$\begin{aligned} \mathcal{L} = & \mu ((\mathbf{w}^+ + \mathbf{w}^-)' \mathbf{j} - 1) + (\mathbf{w}^+ - \mathbf{w}^-)' \mathbf{X} \mathbf{X}' (\mathbf{w}^+ - \mathbf{w}^-) - \beta' \mathbf{e} \\ & - 2(\mathbf{w}^+ - \mathbf{w}^-)' \mathbf{X} \mathbf{K}_b \mathbf{e} + \mathbf{e}' \mathbf{K}_b^2 \mathbf{e} - \alpha^{+'} \mathbf{w}^+ - \alpha^{-'} \mathbf{w}^- + \nu (\mathbf{e}' \mathbf{j} - 1) \end{aligned}$$

where

$$\mu \geq 0, \quad \alpha^+ \geq 0, \quad \alpha^- \geq 0, \quad \beta \geq 0$$

Taking derivatives in respect to  $\mathbf{w}^+$ ,  $\mathbf{w}^-$  and  $\mathbf{e}$ :

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}^+} = -2XK_b\mathbf{e} + \mu\mathbf{j} + 2XX'(\mathbf{w}^+ - \mathbf{w}^-) - \alpha^+ = \mathbf{0}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}^-} = 2XK_b\mathbf{e} + \mu\mathbf{j} - 2XX'(\mathbf{w}^+ - \mathbf{w}^-) - \alpha^- = \mathbf{0}$$

adding the two equations gives  $\alpha^+ = 2\mu\mathbf{j} - \alpha^-$ , implying a lower and upper bound of  $0 \leq \alpha^- \leq 2\mu\mathbf{j}$ .

$$\frac{\partial \mathcal{L}}{\partial \mathbf{e}} = -2K_bX'\mathbf{w} + 2K_b^2\mathbf{e} + \nu\mathbf{j} - \beta$$

- Here  $\mu$  becomes the parameter that can be chosen to control the level of sparsity, while  $\nu$  must then be adapted to ensure  $\mathbf{w}'XX'\mathbf{w} = \mathbf{e}'K_b^2\mathbf{e}$ .
- We initialise with  $\mathbf{e}$  selecting a single document.

# Updates for fixed $\mathbf{e}$

- If we fix  $\mathbf{e}$  and do a coordinate wise ascent solving the quadratic equation exactly the update becomes

$$\mathbf{w}_i \leftarrow \mathbf{w}_i + \frac{1}{4(\mathbf{X}\mathbf{X}')_{ii}} [2\mathbf{X}K_b\mathbf{e} + \mu\mathbf{j} - 2\mathbf{X}\mathbf{X}'\mathbf{w} - \alpha^-]_i.$$

- Once we have iterated to a solution we can then update  $\mathbf{e}$  and  $\nu$  as follows:

$$\begin{aligned}\mathbf{e}_i &\leftarrow \mathbf{e}_i + \frac{1}{4(K_b^2)_{ii}} [2K_b\mathbf{X}'\mathbf{w} - \nu\mathbf{j} - 2K_b^2\mathbf{e} - \beta]_i \\ \nu &\leftarrow \nu + \eta (\mathbf{e}'K_b^2\mathbf{e} - \mathbf{w}'\mathbf{X}\mathbf{X}'\mathbf{w}),\end{aligned}$$

for an appropriate learning rate  $\eta > 0$  with the update to  $\nu$  being applied after each batch of updates.

# Preliminary results

- Just implemented updating of  $\mathbf{w}$  with a fixed (single) document vector  $\mathbf{e}$ .
- Converges very fast because of the sparsity and appears to give very useful summaries of document topic with concise set of words and some negative words.
- Getting cross-lingual retrieval of around 77% on a small dataset for top 5 mate matching.

# Regression Framework for Statistical Machine Translation

- MMR Formulation
- Pre-image Solution
- Kernel Selection
- Experimental Results

- Notation

$\mathcal{X}$  space of source language

$\mathcal{Y}$  space of target language

$\mathcal{H}_x$  Hilbert feature space for source language

$\mathcal{H}_y$  Hilbert feature space for target language

$\phi_x$  mapping  $\mathcal{X} \rightarrow \mathcal{H}_x$

$\phi_y$  mapping  $\mathcal{Y} \rightarrow \mathcal{H}_y$

- Task description:

Based on  $S = \{(x_i, y_i) : x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1, \dots, m\}$ , a set of sample sentence pairs, we are trying to learn a linear function to predict the translation  $y$  for a new source sentence  $x$ , so that

$$\phi_y(y) = \mathbf{W}\phi_x(x) + \mathbf{b}$$

# MMR Formulation

- MMR solves the regression problem by optimizing

- Primal Form:

$$\begin{aligned} \min \quad & \frac{1}{2} \text{tr}(\mathbf{W}^T \mathbf{W}) + C \mathbf{1}^T \boldsymbol{\xi} \\ \text{s.t.} \quad & \langle \phi_y(y_i), \mathbf{W} \phi_x(x_i) + \mathbf{b} \rangle_{\mathcal{H}_y} \geq 1 - \xi_i, \\ & i = 1, \dots, m, \quad \boldsymbol{\xi} \geq \mathbf{0}. \end{aligned}$$

- Dual Form:

$$\begin{aligned} \min \quad & \sum_{i,j=1}^m \alpha_i \alpha_j \kappa_x(x_i, x_j) \kappa_y(y_i, y_j) - \sum_{i=1}^m \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i \phi_y(\mathbf{y}_i) = \mathbf{0}, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m. \end{aligned}$$

- Based on KKT theory,  $\mathbf{W}$  is expressed as

$$\mathbf{W} = \sum_{i=1}^m \alpha_i \phi_y(y_i) \phi_x(x_i)^T$$

# Data Centering & Normalization

- Data centering

$$\begin{cases} \hat{\phi}_x(x) = \phi_x(x) - \frac{1}{m} \sum_{i=1}^m \phi_x(x_i) \\ \hat{\phi}_y(y) = \phi_y(y) - \frac{1}{m} \sum_{i=1}^m \phi_y(y_i) \end{cases}$$

- Remove  $\mathbf{b}$  in the primal form, and the corresponding equality constraints in the dual
- Normalization in L2-norm

$$\begin{cases} \bar{\phi}_x(x) = \hat{\phi}_x(x) / \|\hat{\phi}_x(x)\|_2 \\ \bar{\phi}_y(y) = \hat{\phi}_y(y) / \|\hat{\phi}_y(y)\|_2 \end{cases}$$

- Finally, the prediction is made by

$$\bar{\phi}_y(y) = \sum_{i=1}^m \alpha_i \bar{\phi}_y(y_i) \bar{k}_x(x_i, x)$$

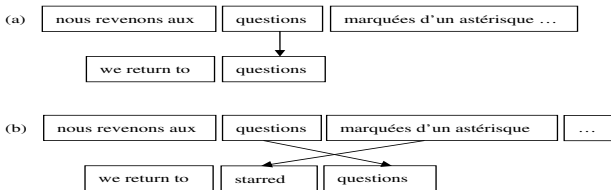
- Pre-image Problem:  $y = \bar{\phi}_y^{-1}(y)$
- It can be achieved by

$$\begin{aligned}y_t &= \arg \max_{y \in \mathcal{Y}(x)} \langle \bar{\phi}_y(y), \mathbf{W} \bar{\phi}_x(x) \rangle_{\mathcal{H}_y} \\ &= \arg \max_{y \in \mathcal{Y}(x)} \sum_{i=1}^m \alpha_i \bar{k}_y(y_i, y) \bar{k}_x(x_i, x)\end{aligned}$$

- Generate a finite set  $\mathcal{Y}(x)$  based on a lexicon
- Dilemma of the "arg max"
  - Enumerate: exponential to the length of  $x$
  - Heuristic search: the inner product cannot be decomposed into a sum of subfunctions to evaluate each component, because of the data centering and normalization

# Pre-image Solution (Cont.)

- Distortion Restriction
  - Only allow adjacent phrases to exchange their positions
  - Example:



- We always have  $x[1 : l_x]$  translated to  $y[1 : l_y]$ , either like (a) or like (b)

- Assumption:

We assume that if  $y$  is a good translation of  $x$ , then  $y[1 : l_y]$  is a good translation of  $x[1 : l_x]$  as well. So we can expect that the inner product  $\langle \bar{\phi}_y(y[1 : l_y]), \mathbf{W}\bar{\phi}_x(x[1 : l_x]) \rangle_{\mathcal{H}_y}$  is large for the hypothesis yielding a good translation.

- Decoding

- Beam search – similar to Pharaoh Decoder
- Limit the derivation of hypotheses as mentioned above
- Score the hypotheses in stacks according to

$$\text{Score}(x[1 : l_x], y[1 : l_y]) = \sum_{i=1}^m \alpha_i \bar{\kappa}_y(y_i, y[1 : l_y]) \bar{\kappa}_x(x_i, x[1 : l_x])$$

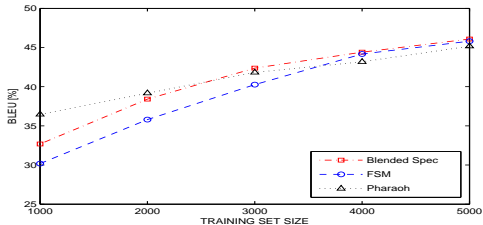
- Blended Spectrum Kernel
  - We count the spectra from unigram up to 7-gram.
  - Spectra are given uniform weight.
- Finite State Machine (FSM) Kernel
  - A set of substrings of sentences are chosen as states.
  - The generation of a sentence is viewed as a finite state automaton, with each word controlling the transition to a new state (the longest one) ending with it.
  - A Fisher kernel for two sentences  $s$  and  $s'$ :  $\kappa(s, s') = U_s^T U_{s'}$ .
  - Component in the Fisher score vector  $U_s$  is indexed by  $(u, a)$ , with the value  $tf((u, a), s) / p_u(a)$ .
  - $tf((u, a), s)$  is the frequency of the transition on word  $a$  from a state  $u$  in sentence  $s$ .
  - To select the state set  $(\Sigma)$ , we keep all the substrings that occur more than twice in a given corpus..
  - $p_u(a)$  is computed as

$$p_u(a) = \sqrt{\frac{f_{u,a} + c}{|A|c + \sum_{a' \in A} f_{u,a'}}$$

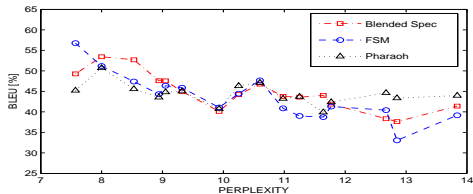
- Baseline System – Pharaoh
  - Translation table size: 10
  - Beam size: 100
  - Developed with minimum error rate training
- Experiments on Biased Corpora
  - Test our models on a corpus with a very small vocabulary (very frequent words only)
  - Filter the French-English portion of the Europal corpus
  - Pick out the sentences of length 5–15, only consisting of words that occur more than 2000 times in the entire corpus
  - 5000 sentences for training, 288 for test, 300 for development

# Experimental Results (Cont.)

- BLEU score over training corpus size

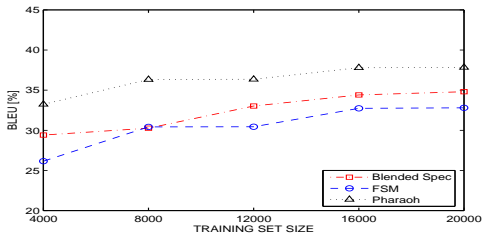


- BLEU score over perplexity
  - Based on 5k training set
  - Repeatedly pick out 100 sentences from the test set randomly



# Experimental Results (Cont.)

- Experiments on regular corpora
  - Sentences of length 5–15 from French-English portion of the Europal corpus
  - 20000 for training, 500 for test, 500 for development
- BLEU score over training set size



- Weighting schemes
  - FSM almost never improved the blended spectrum method
  - $n$ -grams (especially the long ones) are too sparse to give proper probabilities for the weighting
- Primal-dual Sparsity
  - Fully implement
  - How to deflate?
  - Theoretical analysis