Translingual Document Representations from Discriminative Projections

John C. Platt **Kristina Toutanova** **Wen-tau Yih**
Microsoft Research
1 Microsoft Way
Redmond, WA 98005, USA
{jplatt,kristout,scottyih}@microsoft.com

Abstract

Representing documents by vectors that are independent of language enhances machine translation and multilingual text categorization. We use discriminative training to create a projection of documents from multiple languages into a single translingual vector space. We explore two variants to create these projections: Oriented Principal Component Analysis (OPCA) and Coupled Probabilistic Latent Semantic Analysis (CPLSA). Both of these variants start with a basic model of documents (PCA and PLSA). Each model is then made discriminative by encouraging comparable document pairs to have similar vector representations. We evaluate these algorithms on two tasks: parallel document retrieval for Wikipedia and Europarl documents, and cross-lingual text classification on Reuters. The two discriminative variants, OPCA and CPLSA, significantly outperform their corresponding baselines. The largest differences in performance are observed on the task of retrieval when the documents are only comparable and not parallel. The OPCA method is shown to perform best.

1 Introduction

Given the growth of multiple languages on the Internet, Natural Language Processing must operate on dozens of languages. It is becoming critical that computers reach high performance on the following two tasks:

- **Comparable and parallel document retrieval** — Cross-language information retrieval and text categorization have become important with the growth of the Web (Oard and Diekema, 1998). In addition, machine translation (MT) systems can be improved by training on sentences extracted from parallel or comparable documents mined from the Web (Munteanu and Marcu, 2005). Comparable documents can also be used for learning word-level translation lexicons (Fung and Yee, 1998; Rapp, 1999).

- **Cross-language text categorization** — Applications of text categorization, such as sentiment classification (Pang et al., 2002), are now required to run on multiple languages. Categorization is usually trained on the language of the developer: it needs to be easily extended to other languages.

There are two broad approaches to comparable document retrieval and cross-language text categorization. One approach is to translate queries or a training set from different languages into a single target language. Standard monolingual retrieval and classification algorithms can then be applied in the target language.

Alternatively, a cross-language system can project a bag-of-words vector into a translingual lower-dimensional vector space. Ideally, vectors in this space represent the semantics of a document, independent of the language.

The advantage of pre-translation is that MT systems tend to preserve the meaning of documents. However, MT can be very slow (more than 1 second per document), preventing its use on large training sets. When full MT is not practical, a fast word-by-word translation model can be used instead, (Ballesteros and Croft, 1996) but may be less accurate.

Conversely, applying a projection into a low-dimensional space is quick. Linear projection algorithms use matrix-sparse vector multiplication, which can be easily parallelized. However, as seen in section 3, the accuracies of previous projection

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 251–261,
MIT, Massachusetts, USA, 9-11 October 2010.

techniques are not as high as machine translation.

This paper presents two techniques: Oriented PCA and Coupled PLSA. These techniques retain the high speed of projection, while approaching or exceeding the quality level of word glossing. We improve the quality of the projections by the use of discriminative training: we minimize the difference between comparable documents in the projected vector space. Oriented PCA minimizes the difference by modifying the eigensystem of PCA (Diamantaras and Kung, 1996), while Coupled PLSA uses posterior regularization (Graca et al., 2008; Ganchev et al., 2009) on the topic assignments of the comparable documents.

1.1 Previous work

There has been extensive work in projecting monolingual documents into a vector space. The initial algorithm for projecting documents was Latent Semantic Analysis (LSA), which modeled bag-of-word vectors as low-rank Gaussians (Deerwester et al., 1990). Subsequent projection algorithms were based on generative models of individual terms in the documents, including Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

Work on cross-lingual projections followed a similar pattern of moving from Gaussian models to term-wise generative models. Cross-language Latent Semantic Indexing (CL-LSI) (Dumais et al., 1997) applied LSA to concatenated comparable documents from multiple languages. Similarly, Polylingual Topic Models (PLTM) (Mimno et al., 2009) generalized LDA to tuples of documents from multiple languages. The experiments in section 3 use CL-LSI and an algorithm similar to PLTM as benchmarks.

The closest previous work to this paper is the use of Canonical Correlation Analysis (CCA) to find projections for multiple languages whose results are maximally correlated with each other (Vinokourov et al., 2003).

PLSA-, LDA-, and CCA-based cross-lingual models have also been trained without the use of parallel or comparable documents, using only knowledge from a translation dictionary to achieve sharing of topics across languages (Haghighi et al., 2008; Jagarlamudi and Daumé, 2010; Zhang et al., 2010). Such work is complementary to ours and can be used to extend the models to domains lacking parallel documents.

Outside of NLP, researchers have designed algorithms to find discriminative projections. We build on the Oriented Principal Component Analysis (OPCA) algorithm (Diamantaras and Kung, 1996), which finds projections that maximize a signal-to-noise ratio (as defined by the user). OPCA has been used to create discriminative features for audio fingerprinting (Burges et al., 2003).

1.2 Structure of paper

This paper now presents two algorithms for translingual document projection (in section 2): OPCA and Coupled PLSA (CPLSA). To explain OPCA, we first review CL-LSI in section 2.1, then discuss the details of OPCA (section 2.2), and compare it to CCA (section 2.3). To explain CPLSA, we first introduce Joint PLSA (JPLSA), analogous to CL-LSI, in section 2.4, and then describe the details of CPLSA (section 2.5).

We have evaluated these algorithms on two different tasks: comparable document retrieval (section 3.2) and cross-language text categorization (section 3.3). We discuss the findings of the evaluations and extensions to the algorithms in section 4.

2 Algorithms for translingual document projection

2.1 Cross-language Latent Semantic Indexing

Cross-language Latent Semantic Indexing (CL-LSI) is Latent Semantic Analysis (LSA) applied to multiple languages. First, we review the mathematics of LSA.

LSA models an $n \times k$ document-term matrix $\mathbf{D}$, where n is the number of documents and k is the number of terms. The model of the document-term matrix is a low-rank Gaussian. Originally, LSA was presented as performing a Singular Value Decomposition (Deerwester et al., 1990), but here we present it as eigendecomposition, to clarify its relationship with OPCA.

LSA first computes the correlation matrix between terms:

$$\mathbf{C} = \mathbf{D}^T\mathbf{D}. \quad (1)$$

The Rayleigh quotient for a vector $\vec{v}$ with the matrix $\mathbf{C}$ is

$$\frac{\vec{v}^T \mathbf{C} \vec{v}}{\vec{v}^T \vec{v}}, \tag{2}$$

and is equal to the variance of the data projected using the vector $\vec{v}$, normalized by the length of $\vec{v}$, if $\mathbf{D}$ has columns that are zero mean. Good projections retain a large amount of variance. LSA maximizes the Rayleigh ratio by taking its derivative against $\vec{v}$ and setting it to zero. This yields a set of projections that are eigenvectors of $\mathbf{C}$,

$$\mathbf{C}\vec{v}_j = \lambda_j \vec{v}_j, \tag{3}$$

where λ_j is the jth-largest eigenvalue. Each eigenvalue is also the variance of the data when projected by the corresponding eigenvector $\vec{v}_j$. LSA simply uses top d eigenvectors as projections.

LSA is very similar to Principal Components Analysis (PCA). The only difference is that the correlation matrix $\mathbf{C}$ is used, instead of the covariance matrix. In practice, the document-term matrix $\mathbf{D}$ is sparse, so the column means are close to zero, and the correlation matrix is close to the covariance matrix.

There are a number of methods to form the document-term matrix $\mathbf{D}$. One method that works well in practice is to compute the log(tf)-idf weighting: (Dumais, 1990; Wild et al., 2005)

$$D_{ij} = \log_2(f_{ij} + 1) \log_2(n/d_j), \tag{4}$$

where f_{ij} is the number of times term j occurs in document i, n is the total number of documents, and d_j is the total number of documents that contain term j. Applying a logarthm to the term counts makes the distribution of matrix entries approach Gaussian, which makes the LSA model more valid.

Cross-language LSI is an application of LSA where each row of $\mathbf{D}$ is formed by concatenating comparable or parallel documents in multiple languages. If a single term occurs in multiple languages, the term only has one slot in the concatenation, and the term count accumulates for all languages. Such terms could be proper nouns, such as "Smith" or "Merkel".

In general, the elements of $\mathbf{D}$ are computed via

$$D_{ij} = \log_2 \left(\sum_m f_{ij}^m + 1 \right) \log_2(n/d_j), \tag{5}$$

where f_{ij}^m is the number of times term j occurs in document i in language m. Here, d_j is the number of documents term j appears in, and n is the total number of documents across all languages.

Because CL-LSI is simply LSA applied to concatenated documents, it models terms in document vectors jointly across languages as a single low-rank Gaussian.

2.2 Oriented Principal Component Analysis

The limitations of CL-LSI can be illustrated by considering Oriented Principal Components Analysis (OPCA), a generalization of PCA. A user of OPCA computes a signal covariance matrix $\mathbf{S}$ and a noise covariance matrix $\mathbf{N}$. OPCA projections $\vec{v}_j$ maximize the ratio of the variance of the signal projected by $\vec{v}_j$ to the variance of the noise projected by $\vec{v}_j$. This signal-to-noise ratio is the generalized Rayleigh quotient: (Diamantaras and Kung, 1996)

$$\frac{\vec{v}^T \mathbf{S} \vec{v}}{\vec{v}^T \mathbf{N} \vec{v}}. \tag{6}$$

Taking the derivative of the Rayleigh quotient with respect to the projections $\vec{v}$ and setting it to zero yields the generalized eigenproblem

$$\mathbf{S}\vec{v}_j = \lambda_j \mathbf{N}\vec{v}_j. \tag{7}$$

This eigenproblem has no local minima, and can be solved with commonly available parallel code.

PCA is a specialization of OPCA, where the noise covariance matrix is assumed to be the identity (i.e., uncorrelated noise). PCA projections maximize the signal-to-noise ratio where the signal is the empirical covariance of the data, and the noise is spherical white noise. PCA projections are not truly appropriate for forming multilingual document projections.

Instead, we want multilingual document projections to maximize the projected covariance of document vectors across all languages, while simultaneously minimizing the projected distance between comparable documents (see Figure 1). OPCA gives us a framework for finding such discriminative projections. The covariance matrix for all documents is the signal covariance in OPCA, and captures the meaning of documents across all languages. The projection of this covariance matrix should be maximized. The covariance matrix formed from differences between comparable documents is the noise

covariance in OPCA: we wish to minimize the latter covariance, to make the projection language-independent.

Specifically, we create the weighted document-term matrix $\mathbf{D}_m$ for each language:

$$D_{ij,m} = \log_2(f_{ij}^m + 1) log_2(n/d_j). \quad (8)$$

We then derive a signal covariance matrix over all languages:

$$\mathbf{S} = \sum_m \mathbf{D}_m^T \mathbf{D}_m / n - \vec{\mu}_m^T \vec{\mu}_m, \quad (9)$$

where $\vec{\mu}_m$ is the mean of each $\mathbf{D}_m$ over its columns, and a noise covariance matrix,

$$\mathbf{N} = \sum_m (\mathbf{D}_m - \mathbf{D})^T (\mathbf{D}_m - \mathbf{D})/n + \gamma \mathbf{I}, \quad (10)$$

where $\mathbf{D}$ is the mean across all languages of the document-term matrix,

$$\mathbf{D} = \frac{1}{M} \sum_m \mathbf{D}_m, \quad (11)$$

and M is the number of languages. Applying equation (7) to these matrices and taking the top generalized eigenvectors yields the projection matrix for OPCA.

Note the regularization term of $\gamma \mathbf{I}$ in equation (10). The empirical sample of comparable documents may not cover the entire space of translation noise the system will encounter in the test set. For safety, we add a regularizer that prevents the variance of a term from getting too small. We tuned γ on the development sets in section 3.2: for log(tf)-idf weighted vectors, $C = 0.1$ works well for the data sets and dimensionalities that we tried. We use $C = 0.1$ for all final tests.

2.3 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a technique that is related to OPCA. CCA was kernelized and applied to creating cross-language document models by (Vinokourov et al., 2003). In CCA, a linear projection is found for each language, such that the projections of the corpus from each language are maximally correlated with each other. Similar to OPCA, this linear projection can be found by finding the top generalized eigenvectors of the system

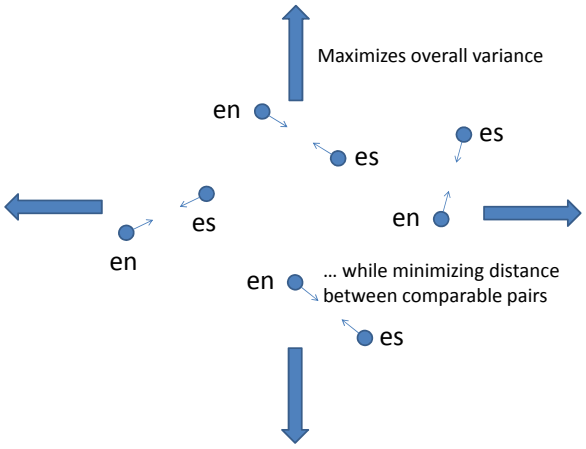

Figure 1: OPCA finds a projection that maximizes the variance of all documents, while minimizing distance between comparable documents

(7), where $\mathbf{S}$ is now a matrix of cross-correlations that the projection maximizes,

$$\mathbf{S} = \begin{bmatrix} 0 & \mathbf{C}_{12} \\ \mathbf{C}_{21} & 0 \end{bmatrix}, \quad (12)$$

and $\mathbf{N}$ is a matrix of autocorrelations that the projection minimizes,

$$\mathbf{N} = \begin{bmatrix} \mathbf{C}_{11} + \gamma I & 0 \\ 0 & \mathbf{C}_{22} + \gamma I \end{bmatrix}. \quad (13)$$

Here, C_{ij} is the (cross-)covariance matrix, with dimension equal to the vocabulary size, that is computed between the document vectors for languages i and j. Analogous to OPCA, γ is a regularization term, set by optimizing performance on a validation set. Like OPCA, these matrices can be generalized to more than two languages. Unlike OPCA, CCA finds projections that maximize the cross-covariance between the projected vectors, instead of minimizing Euclidean distance.[1]

By definition, CCA cannot take advantage of the information that same term occurs simultaneously in comparable documents. As shown in section 3, this

[1]Note that the eigenvectors have length equal to the sum of the length of the vocabularies of each language. The projections for each language are created by splitting the eigenvectors into sections, each with length equal to the vocabulary size for each language.

information is useful and helps OPCA perform better then CCA. In addition, CCA encourages comparable documents to be projected to vectors that are mutually linearly predictable. This is not the same OPCA's projected vectors that have low Euclidean distance: the latter may be preferred by algorithms that consume the projections.

2.4 Cross-language Topic Models

We now turn to a baseline generative model that is analogous to CL-LSI. Our baseline joint PLSA model (JPLSA) is closely related to the poly-lingual LDA model of (Mimno et al., 2009). The graphical model for JPLSA is shown at the top in Figure 2. We describe the model for two languages, but it is straightforward to generalize to more than two languages, as in (Mimno et al., 2009).

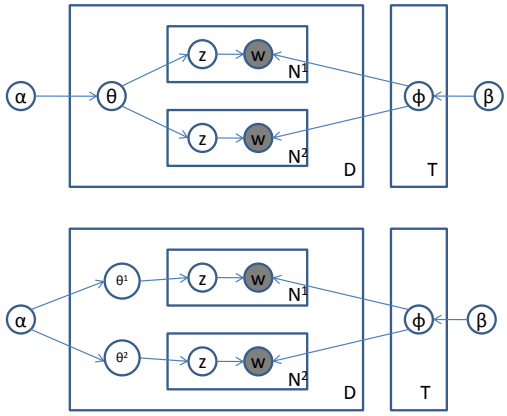

Figure 2: Graphical models for JPLSA (top) and CPLSA (bottom)

The model sees documents d_i as sequences of words $w_1, w_2, \ldots, w_{n_i}$ from a vocabulary V. There are T cross-language topics, each of which has a distribution ϕ_t over words in V. In the case of models for two languages, we define the vocabulary V to contain word types from both languages. In this way, each topic is shared across languages.

Each topic-specific distribution ϕ_t, for $t = 1 \ldots T$, is drawn from a symmetric Dirichlet prior with concentration parameter β. Given the topic-specific word distributions, the generative process for a corpus of paired documents $[d_i^1, d_i^2]$ in two languages L_1 and L_2 is described in the next paragraph.

For each pair of documents, pick a distribution over topics θ_i, from a symmetric Dirichlet prior with concentration parameter α. Then generate the documents d_i^1 and d_i^2 in turn. Each word token in each document is generated independently by first picking a topic z from a multinomial distribution with parameter θ_i (MULTI(θ_i)), and then generating the word token from the topic-specific word distribution for the chosen topic MULTI(ϕ_z).

The probability of a document pair $[d^1, d^2]$ with words $[w_1^1, w_2^1, \ldots, w_{n_1}^1]$, $[w_1^2, w_2^2, \ldots, w_{n_2}^2]$, topic assignments $[z_1^1, \ldots, z_{n_1}^1]$, $[z_1^2, \ldots, z_{n_2}^2]$, and a common topic vector θ is given by:

$$P(\theta|\alpha) \prod_{j=1}^{n_1} P(z_j^1|\theta) P(w_j^1|\phi_{z_j^1}) \prod_{j=1}^{n_2} P(z_j^2|\theta) P(w_j^2|\phi_{z_j^2})$$

The difference between the JPLSA model and the poly-lingual topic model of (Mimno et al., 2009) is that we merge the vocabularies in the two languages and learn topic-specific word distributions over these merged vocabularies, instead of having pairs of topic-specific word distributions, one for each language, like in (Mimno et al., 2009). Thus our model is more similar to the CL-LSI model, because it can be seen as viewing a pair of documents in two languages as one bigger document containing the words in both documents.

Another difference between our model and the poly-lingual LDA model of (Mimno et al., 2009) is that we use maximum aposteriori (MAP) instead of Bayesian inference. Recently, MAP inference was shown to perform comparably to the best inference method for LDA (Asuncion et al., 2009), if the hyper-parameters are chosen optimally for the inference method. Our initial experiments with Bayesian versus MAP inference for parallel document retrieval using JPLSA confirmed this result. In practice our baseline model outperforms poly-lingual LDA as mentioned in our experiments.

2.5 Coupled Probabilistic Latent Semantic Analysis

The JPLSA model assumes that a pair of translated or comparable documents have a common topic distribution θ. JPLSA fits its parameters to optimize the probability of the data, given this assumption.

For the task of comparable document retrieval, we want our topic model to assign similar topic distributions θ to a pair of corresponding documents. But

this is not exactly what the JPLSA model is doing. Instead, it derives a common topic vector θ which explains the union of all tokens in the English and foreign documents, instead of making sure that the best topic assignment for the English document is close to the best topic assignment of the foreign document. This difference becomes especially apparent when corresponding documents have different lengths. In this case, the model will tend to derive a topic vector θ which explains the longer document best, making the sum of the two documents' log-likelihoods higher. Modeling the shorter document's best topic carries little weight.

Modeling both documents equally is what Coupled PLSA (CPLSA) is designed to do. The graphical model for CPLSA is shown at the bottom of Figure 2. In this figure, the topic vectors of a pair of documents in two languages are shown completely independent. We use the log-likelihood according to this model, but also add a regularization term, which tries to make the topic assignments of corresponding documents close. In particular, we use posterior regularization (Graca et al., 2008; Ganchev et al., 2009) to place linear constraints on the expectations of topic assignments to two corresponding documents.

For two linked documents d_1 and d_2, we would like our model to be such that the expected fraction of tokens in d_1 that get assigned topic t is approximately the same as the expected fraction of tokens in d_2 that get assigned the same topic t, for each topic $t = 1 \dots T$. This is exactly what we need to make each pair of corresponding documents close.

Let $\mathbf{z^1}$ and $\mathbf{z^2}$ denote vectors of topic assignments to the tokens in document d^1 and d^2, respectively. Their dimensionality is equal to the lengths of the two documents, n_1 and n_2. We define a space of posterior distributions Q over hidden topic assignments to the tokens in d^1 and d^2, that has the desired property: the expected fraction of each topic is approximately equal in d^1 and d^2. We can formulate this constrained space Q as follows:

$$Q = \{q_1(\mathbf{z^1}), q_2(\mathbf{z^2})\}$$

such that

$$\mathbf{E}_{q_1}[\frac{\sum_{j=1}^{n_1} \mathbf{1}(z_j^1 = t)}{n_1}] - \mathbf{E}_{q_2}[\frac{\sum_{j=1}^{n_2} \mathbf{1}(z_j^2 = t)}{n_2}] \leq \epsilon_t$$

$$\mathbf{E}_{q_2}[\frac{\sum_{j=1}^{n_2} \mathbf{1}(z_j^2 = t)}{n_2}] - \mathbf{E}_{q_1}[\frac{\sum_{j=1}^{n_1} \mathbf{1}(z_j^1 = t)}{n_1}] \leq \epsilon_t$$

We then formulate an objective function that maximizes the log-likelihood of the data while simultaneously minimizing the KL-divergence between the desired distribution set Q and the posterior distribution according to the model: $P(\mathbf{z_1}|d_1, \theta_1, \phi)$ and $P(\mathbf{z_2}|d_2, \theta_2, \phi)$.

The objective function for a single document pair is as follows:

$$\begin{aligned}\log P(d_1|\theta_1, \phi) + \log P(d_2|\theta_2, \phi) \\ -\mathbf{KL}(Q||P(\mathbf{z_1}|d_1, \theta_1, \phi), P(\mathbf{z_2}|d_2, \theta_2, \phi)) \\ -||\epsilon||\end{aligned}$$

The final corpus-wide objective is summed over document-pairs, and also contains terms for the probabilities of the parameters θ and ϕ given the Dirichlet priors. The norm of ϵ is minimized, which makes the expected proportions of topics in two documents as close as possible.

Following (Ganchev et al., 2009), we fit the parameters by an EM-like algorithm, where for each document pair, after finding the posterior distribution of the hidden variables, we find the KL-projection of this posterior onto the constraint set, and take expected counts with respect to this projection; these expected counts are used in the M-step. The projection is found using a simple projected gradient algorithm.[2]

For both the baseline JPLSA and the CPLSA models, we performed learning through MAP inference using EM (with a projection step for CPLSA). We did up to 500 iterations for each model, and did early stopping based on task performance on the development set. The JPLSA model required more iterations before reaching its peak accuracy, tending to require around 300 to 450 iterations for convergence. CPLSA required fewer iterations, but each iteration was slower due to the projection step.

[2]We initialized the models deterministically by assigning each word to exactly one topic to begin with, such that all topics have roughly the same number of words. Words were sorted by frequency and thus words of similar frequency are more likely to be assigned to the same topic.This initialization method outperformed random initialization and we use it for all models.

All models use $\alpha = 1.1$ and $\beta = 1.01$ for the values of the concentration parameters. We found that the performance of the models was not very sensitive to these values, in the region that we tested ($\alpha, \beta \in [1.001, 1.1]$). Higher hyper-parameter values resulted in faster convergence, but the final performance was similar across these different values.

3 Experimental validation

We test the proposed discriminative projections versus more established cross-language models on the two tasks described in the introduction: retrieving comparable documents from a corpus, and training a classifier in one language and using it in another. We measure accuracy on a test set, and also examine the sensitivity to dimensionality of the projection on development sets.

3.1 Speed of training and evaluation

We first test the speed of the various algorithms discussed in this paper, compared to a full machine translation system. When finding document projections, CL-LSI, OPCA, CCA, JPLSA, and CPLSA are equally fast: they perform a matrix multiplication and require $O(nk)$ operations, where n is the number of distinct words in the documents and k is the dimensionality of the projection.[3] A single CPU core can read the indexed documents into memory and take logarithms at 216K words per second. Projecting into a 2000-dimensional space operates at 41K words per second. Translating word-by-word operates at 274K words per second. In contrast, machine translation processes 50 words per second, approximately 3 orders of magnitude slower.

Total training time for OPCA on 43,380 pairs of comparable documents was 90 minutes, running on an 8-core CPU for 2000 dimensions. On the same corpus, JPLSA requires 31 minutes per iteration and CPLSA requires 377 minutes per iteration. CPLSA requires a factor of five times fewer iterations: overall, it is twice as slow as JPLSA.

3.2 Retrieval of comparable documents

In comparable document retrieval, a query is a document in one language, which is compared to a corpus of documents in another language. By mapping all documents into the same vector space, the comparison is a vector comparison. For our experiments with CL-LSI, OPCA, and CCA, we use cosine similarity between vectors to rank the documents.

For the JPLSA and CPLSA models, we map the documents to corresponding topic vectors θ, and compute distance between these probability vectors. The mapping to topic vectors requires EM iterations, or folding-in (Hofmann, 1999). We found that performing a single EM iteration resulted in best performance so we used this for all models. For computing distance we used the L1-norm of the difference, which worked a bit better than the Jensen-Shannon divergence between the topic vectors used in (Mimno et al., 2009).

We test all algorithms on the Europarl data set of documents in English and Spanish, and a set of Wikipedia articles in English and Spanish that contain interlanguage links between them (i.e., articles that the Wikipedia community have identified as comparable across languages).

For the Europarl data set, we use 52,685 documents as training, 11,933 documents as a development set, and 18,415 documents as a final test set. Documents are defined as speeches by a single speaker, as in (Mimno et al., 2009).[4] For the Wikipedia set, we use 43,380 training documents, 8,675 development documents, and 8,675 final test set documents.

For both corpora, the terms are extracted by word-breaking all documents, removing the top 50 most frequent terms and keeping the next 20,000 most frequent terms. No stemming or folding is applied.

We assess performance by testing each document in English against all possible documents in Spanish, and *vice versa*. We measure the Top-1 accuracy (i.e., whether the true comparable is the closest in the test set), and the Mean Reciprocal Rank of the true comparable, and report the average performance over the two retrieval directions. Ties are counted as errors.

We tuned the dimensionality of the projections on the development set, as shown in Figures 3 and 4.

[3] For JPLSA and CPLSA this is the case only when performing a single EM iteration at test time, which we found to perform best.

[4] The training section contains documents from the years 96 through 99 and the year 02; the dev section contains documents from 01, and the test section contains documents from 00 plus the first 9 months of 03.

We chose the best dimension on the development set for each algorithm, and used it on the final test set. The regularization γ was tuned for CCA: $\gamma = 10$ for Europarl, and $\gamma = 3$ for Wikipedia.

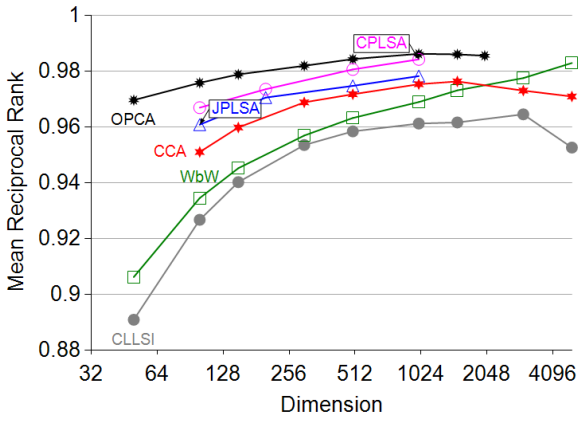

Figure 3: Mean reciprocal rank versus dimension for Europarl

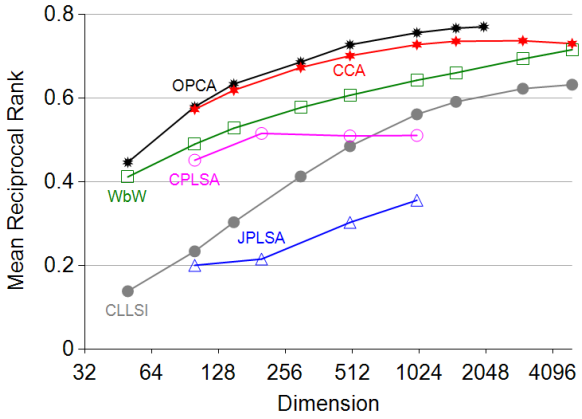

Figure 4: Mean reciprocal rank versus dimension for Wikipedia

In the two figures, we evaluate the five projection methods, as well as a word-by-word translation method (denoted by WbW in the graphs). Here "word-by-word" refers to using cosine distance after applying a word-by-word translation model to the Spanish documents.

The word-by-word translation model was trained on the Europarl training set, using the WDHMM model (He, 2007), which performs similarly to IBM Model 4. The probability matrix of generating English words from Spanish words was multiplied by each document's log(tf)-idf vector to produce a translated document vector. We found that multiplying the probability matrix to the log(tf)-idf vector was more accurate on the development set than multiplying the tf vector directly. This vector was either tested as-is, or mapped through LSA learned from the English training set of the corpus. In the figures, the dimensionality of WbW translation refers to the dimensionality of monolingual LSA.

The overall ordering of the six models is different for the Europarl and Wikipedia development datasets. The discriminative models outperform the corresponding generative ones (OPCA vs CLLSI) and (CPLSA vs JPLSA) for both datasets, and OPCA performs best overall, dominating the best fast-translation based model, as well as the other projection methods, including CCA.

On Europarl, JPLSA and CPLSA outperform CLLSI, with the best dimension or JPLSA also slightly outperforming the best setting for the word-by-word translation model, whereas on Wikipedia the PLSA-based models are significantly worse than the other models.

The results on the final test set, evaluating each model using its best dimensionality setting, confirm the trends observed on the development set. The final results are shown in Tables 1 and 2. For these experiments, we use the unpaired t-test with Bonferroni correction to determine the smallest set of algorithms that have statistically significantly better accuracy than the rest. The p-value threshold for significance is chosen to be 0.05. The accuracies for these significantly superior algorithms are shown in boldface.

For Wikipedia and Europarl, we include an additional baseline model,"Untranslated": this refers to applying cosine distance to both the Spanish and English documents directly (since they share some vocabulary terms). For Wikipedia, comparable documents seem to share many common terms, so cosine distance between untranslated documents is a reasonable benchmark.

From the final Europarl results we can see that the best models can learn to retrieve parallel documents from the narrow Europarl domain very well. All dimensionality reduction methods can learn from

cleanly parallel data, but discriminative training can bring additional error reduction.

In previously reported work, (Mimno et al., 2009) evaluate parallel document retrieval using PLTM on Europarl speeches in English and Spanish, using training and test sets of size similar to ours. They report an accuracy of 81.2% when restricting to test documents of length at least 100 and using 50 topics. JPLSA with 50 topics obtains accuracy of 98.9% for documents of that length.

The final Wikipedia results are also similar to the the development set results. The problem setting for Wikipedia is different, because corresponding documents linked in Wikipedia may have widely varying degrees of parallelism. While most linked documents share some main topics, they could cover different numbers of sub-topics at varying depths. Thus the training data of linked documents is noisy, which makes it hard for projection methods to learn. The word-by-word translation model in this setting is trained on clean, but out-of-domain parallel data (Europarl), so it has the disadvantage that it may not have a good coverage of the vocabulary; however, it is not able to make use of the Wikipedia training data since it requires sentence-aligned translations. We find it encouraging that the best projection method OPCA outperformed word-by-word translation. This means that OPCA is able to uncover topic correspondence given only comparable document pairs, and to learn well in this noisy setting.

The PLSA-based models fare worse on Wikipedia document retrieval. CPLSA outperforms JPLSA more strongly, but both are worse than CL-LSI and even the Untranslated baseline. We think this is partly explained by the diverse vocabulary in the heterogenous Wikipedia collection. All other models use log(tf)-idf weighting, which automatically assigns importance weights to terms, whereas the topic models use word counts. This weighting is very useful for Wikipedia. For example, if we apply the untranslated matching using raw word counts, the MRR is 0.1024 on the test set, compared to 0.5383 for log(tf)-idf. We hypothesize that using a hierarchical topic model that automatically learns about more general and more topic-specific words would be helpful in this case. It is also possible that PLSA-based models require cleaner data to learn well.

The overall conclusion is that OPCA outperformed all other document retrieval methods we tested, including fast machine translation of documents. Additionally, both discriminative projection methods outperformed their generative counterparts.

Algorithm	Dimension	Accuracy	MRR
OPCA	1000	**0.9742**	0.9806
CPLSA	1000	**0.9716**	0.9782
Word-by-word	N/A	**0.9707**	0.9779
Word-by-word	5000	**0.9706**	0.9778
JPLSA	1000	0.9645	0.9726
CCA	1500	0.9613	0.9705
CL-LSI	3000	0.9457	0.9595
Untranslated	N/A	0.1595	0.2564

Table 1: Test results for comparable document retrieval in Europarl. Boldface indicates statistically significant superior results.

Algorithm	Dimension	Accuracy	MRR
OPCA	2000	**0.7255**	0.7734
Word-by-word	N/A	0.7033	0.7467
CCA	1500	0.6894	0.7378
Word-by-word	5000	0.6786	0.7236
CL-LSI	5000	0.5302	0.6130
Untranslated	N/A	0.4692	0.5383
CPLSA	200	0.4579	0.5130
JPLSA	1000	0.3322	0.3619

Table 2: Test results for comparable document retrieval in Wikipedia. Boldface indicates statistically significant best result.

3.3 Cross-language text classification

The second task is to train a text categorization system in one language, and test it with documents in another. To evaluate on this task, we use the Multilingual Reuters Collection, defined and provided by (Amini et al., 2009). We test the English/Spanish language pair. The collection has news articles in English and Spanish, each of which has been translated to the other by the Portage translation system (Ueffing et al., 2007).

From the English news corpus, we take 13,131 documents as training, 1,875 documents as development, and 1,875 documents as test. We take the English training documents translated into Spanish as our comparable training data. For testing, we use the entire Spanish news corpus of 12,342 documents, ei-

ther mapped with cross-lingual projection, or translated by Portage.

The data set was provided by (Amini et al., 2009) as already-processed document vectors, using BM25 weighting. Thus, we only test OPCA, CL-LSI, and related methods: JPLSA and CPLSA require modeling the term counts directly.

The performance on the task is measured by classification accuracy on the six disjoint category labels defined by (Amini et al., 2009). To introduce minimal bias due to the classifier model, we use 1-nearest neighbor on top of the cosine distance between vectors as a classifier. For all of the techniques, we treated the vocabulary in each language as completely separate, using the top 10,000 terms from each language.

Note that no Spanish labeled data is provided for training any of these algorithms: only English and translated English news is labeled. The optimal dimension (and γ for CCA) on the development set was chosen to maximize the accuracy of English classification and translated English-to-Spanish classification.

Algorithm	Dim.	English Accuracy	Spanish Accuracy
Full MT	50	**0.8483**	**0.6484**
OPCA	100	**0.8412**	0.5954
Word-by-word	50	**0.8483**	0.5780
CCA	150	**0.8388**	0.5384
Full MT	N/A	0.8046	0.5323
CL-LSI	150	**0.8401**	0.5105
Word-by-word	N/A	0.8046	0.4481

Table 3: Test results for cross-language text categorization

The test classification accuracy is shown in Table 3. As above, the smallest set of superior algorithms as determined by Bonferroni-corrected t-tests are shown in boldface. The results for MT and word-by-word translation use the log(tf)-idf vector directly for documents that were written in English, and use a Spanish-to-English translated vector if the document was written in Spanish. As in section 3.2, word-by-word translation multiplied each log(tf)-idf vector by the translation probability matrix trained on Europarl.

The tests show that OPCA is better than CCA, CL-LSI, plain word-by-word translation, and even full translation for Spanish documents. However, if we post-process full translation by an LSI model trained on the English training set, full translation is the most accurate. If full translation is time-prohibitive, then OPCA is the best method: it is significantly better than word-by-word translation followed by LSI.

4 Discussion and Extensions

OPCA extends naturally to multiple languages. However, it requires memory and computation time that scales quadratically with the size of the vocabulary. As the number of languages goes up, it may become impractical to perform OPCA directly on a large vocabulary.

Researchers have solved the problem of scaling OPCA by using Distortion Discriminant Analysis (DDA) (Burges et al., 2003). DDA performs OPCA in two stages which avoids the need for solving a very large generalized eigensystem. As future work, DDA could be applied to mapping documents in many languages simultaneously.

Spherical Admixture Models (Reisinger et al., 2010) have recently been proposed that combine an LDA-like hierarchical generative model with the use of tf-idf representations. A similar model could be used for CPLSA: future work will show whether such a model can outperform OPCA.

5 Conclusions

This paper presents two different methods for creating discriminative projections: OPCA and CPLSA. Both of these methods avoid the use of artificial concatenated documents. Instead, they model documents in multiple languages, with the constraint that comparable documents should map to similar locations in the projected space.

When compared to other techniques, OPCA had the highest accuracy while still having a run-time that allowed scaling to large data sets. We therefore recommend the use of OPCA as a pre-processing step for large-scale comparable document retrieval or cross-language text categorization.

References

Massih-Reza Amini, Nicolas Usunier, and Cyril Goutte. 2009. Learning from multiple partially observed views - an application to multilingual text categorization. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, pages 28–36.

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 27–34.

Lisa Ballesteros and Bruce Croft. 1996. Dictionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, pages 791–801.

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Christopher J.C. Burges, John C. Platt, and Soumya Jana. 2003. Distortion discriminant analysis for audio fingerprinting. *IEEE Transactions on Speech and Audio Processing*, 11(3):165–174.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Konstantinos I. Diamantaras and S.Y. Kung. 1996. *Principal Component Neural Networks: Theory and Applications*. Wiley-Interscience.

Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval*.

Susan T. Dumais. 1990. Enhancing performance in latent semantic indexing (LSI) retrieval. Technical Report TM-ARH-017527, Bellcore.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of COLING-ACL*, pages 414–420.

Kuzman Ganchev, Joao Graca, Jennifer Gillenwater, and Ben Taskar. 2009. Posterior regularization for structured latent variable models. Technical Report MS-CIS-09-16, University of Pennsylvania.

Joao Graca, Kuzman Ganchev, and Ben Taskar. 2008. Expectation maximization and posterior constraints. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press, Cambridge, MA.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proc. ACL*, pages 771–779.

Xiaodong He. 2007. Using word-dependent transition models in HMM based word alignment for statistical machine translation. In *ACL 2nd Statistical MT workshop*, pages 80–87.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence*, pages 289–296.

Jagadeesh Jagarlamudi and Hal Daumé, III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *ECIR*.

David Mimno, Hanna W. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 880–889.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504.

Douglas W. Oard and Anne R. Diekema. 1998. Cross-language information retrieval. In Martha Williams, editor, *Annual Review of Information Science (ARIST)*, volume 33, pages 223–256.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proc. EMNLP*, pages 79–86.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the ACL*, pages 519–526.

Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J. Mooney. 2010. Spherical topic models. In *Proc. ICML*.

Nicola Ueffing, Michel Simard, Samuel Larkin, and J. Howard Johnson. 2007. NRC's PORTAGE system for WMT 2007. In *ACL-2007 2nd Workshop on SMT*, pages 185–188.

Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. 2003. Inferring a semantic representation of text via cross-language correlation analysis. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1473–1480, Cambridge, MA. MIT Press.

Fridolin Wild, Christina Stahl, Gerald Stermsek, and Gustaf Neumann. 2005. Parameters driving effectiveness of automated essay scoring with LSA. In *Proceedings 9th Internaional Computer-Assisted Assessment Conference*, pages 485–494.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proc. ACL*, pages 1128–1137, Uppsala, Sweden. Association for Computational Linguistics.