Improving Bilingual Projections via Sparse Covariance Matrices

Jagadeesh Jagarlamudi
University of Maryland
College Park, USA
jags@umiacs.umd.edu

Raghavendra Udupa
Microsoft Research
Bangalore, India
raghavu@microsoft.com

Hal Daumé III
University of Maryland
College Park, USA
hal@umiacs.umd.edu

Abhijit Bhole
Microsoft Research
Bangalore, India
v-abbhol@microsoft.com

Abstract

Mapping documents into an interlingual representation can help bridge the language barrier of cross-lingual corpora. Many existing approaches are based on word co-occurrences extracted from aligned training data, represented as a covariance matrix. In theory, such a covariance matrix should represent semantic equivalence, and *should* be highly sparse. Unfortunately, the presence of noise leads to dense covariance matrices which in turn leads to suboptimal document representations. In this paper, we explore techniques to recover the desired sparsity in covariance matrices in two ways. First, we explore word association measures and bilingual dictionaries to weigh the word pairs. Later, we explore different selection strategies to remove the noisy pairs based on the association scores. Our experimental results on the task of aligning comparable documents shows the efficacy of sparse covariance matrices on two data sets from two different language pairs.

1 Introduction

Aligning documents from different languages arises in a range of tasks such as parallel phrase extraction (Gale and Church, 1991; Rapp, 1999), mining translations for out-of-vocabulary words for statistical machine translation (Daume III and Jagarlamudi, 2011) and document retrieval (Ballesteros and Croft, 1996; Munteanu and Marcu, 2005). In this task, we are given a comparable corpora and some documents in one language are assumed to have a comparable document in the other language and the goal is to recover this hidden alignment. In this paper, we address this problem by mapping the documents into a common subspace (interlingual representation). This common subspace generalizes the notion of vector space model for cross-lingual applications (Turney and Pantel, 2010).

Most of the existing approaches use manually aligned document pairs to find a common subspace in which the aligned document pairs are maximally correlated. The sub-space can be found using either generative approaches based on topic modeling (Mimno et al., 2009; Jagarlamudi and Daumé III, 2010; Zhang et al., 2010; Vu et al., 2009) or discriminative approaches based on variants of Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) (Susan T. Dumais, 1996; Vinokourov et al., 2003; Platt et al., 2010; Haghighi et al., 2008). Both styles rely on document level term co-occurrences to find the latent representation.

The discriminative approaches capture essential word co-occurrences in terms of two monolingual covariance matrices and a cross-covariance matrix. Subsequently, they use these covariance matrices to find projection directions in each language such that aligned documents lie close to each other (Sec. 2). The strong reliance of these approaches on the covariance matrices leads to problems, especially with the noisy data caused either by the noisy words in a document or the noisy document alignments. Noisy data is not uncommon and is usually the case with data collected from community based resources such as Wikipedia. This degrades performance of a

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 930–940,
Edinburgh, Scotland, UK, July 27–31, 2011.

variety of tasks, such as transliteration Mining (Klementiev and Roth, 2006; Hermjakob et al., 2008; Ravi and Knight, 2009) and multilingual web search (Gao et al., 2009).

In this paper, we address the problem of identifying and removing noisy entries in the covariance matrices. We address this problem in two stages. In the first stage, we explore the use of word association measures such as Mutual Information (MI) and Yule's ω (Reis and Judd, 2000) in computing the strength of a word pair (Sec. 3.1). We also explore the use of bilingual dictionaries developed from cleaner resources such as parallel data. In the second stage, we use the association strengths in filtering out the noisy word pairs from the covariance matrices. We pose this as a word pair selection problem and explore multiple strategies (Sec. 3.2).

We evaluate the utility of sparse covariance matrices in improving the bilingual projections incrementally (Sec. 4). We first report results on synthetic multi-view data where the true correspondences between features of different views are available. Moreover, this also lets us systematically explore the effect of noise level on the accuracy. Our experimental results show a significant improvement when the true correspondences are available. Later, we report our experimental results on the document alignment task on Europarl and Wikipedia data sets and on two language pairs. We found that sparsifying the covariance matrices helps in general, but using cleaner resource such bilingual dictionaries performed best.

2 Canonical Correlation Analysis (CCA)

In this section, we describe how Canonical Correlation Analysis is used to solve the problem of aligning bilingual documents. We mainly focus on representing the solution of CCA in terms of covariance matrices. Since most of the existing discriminative approaches are variants of CCA, showing the advantage of recovering sparseness in CCA makes it applicable to the other variants as well.

Given a training data of n aligned document pairs, CCA finds projection directions for each language, so that the documents when projected along these directions are maximally correlated (Hotelling, 1936). Let X ($d_1 \times n$) and Y ($d_2 \times n$) be the representation of data in both the languages and further assume that the data is centered (subtract the mean vector from each document *i.e.* $\mathbf{x}_i \leftarrow \mathbf{x}_i - \mu_x$ and $\mathbf{y}_i \leftarrow \mathbf{y}_i - \mu_y$). Then CCA finds projection directions $\mathbf{a}$ and $\mathbf{b}$ which maximize:

$$\frac{\mathbf{a}^T X Y^T \mathbf{b}}{\sqrt{\mathbf{a}^T X X^T \mathbf{a}}\sqrt{\mathbf{b}^T Y Y^T \mathbf{b}}}$$
$$\text{s.t. } \mathbf{a}^T X X^T \mathbf{a} = 1 \ \& \ \mathbf{b}^T Y Y^T \mathbf{b} = 1$$

The projection directions are obtained by solving the generalized eigen system:

$$\begin{bmatrix} 0 & C^{xy} \\ C^{yx} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} (1\text{-}\lambda)C^{xx}+\lambda I & 0 \\ 0 & (1\text{-}\lambda)C^{yy}+\lambda I \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \quad (1)$$

where $C^{xx} = XX^T$, $C^{yy} = YY^T$ are the monolingual covariance matrices, $C^{xy} = XY^T$ is the cross-covariance matrix and λ is the regularization parameter. Using these eigenvectors as columns, we form the projection matrices A and B. These projection matrices are used to map documents in both the languages into interlingual representation.

Given any new pair of documents, their similarity is computed by first mapping them into the lower dimensions space and computing the cosine similarity between their projections. In general, using all the eigenvectors is sub optimal and thus retaining top eigenvectors leads to better generalizability.

3 Covariance Selection

As shown above, the underlying objective function in most of the discriminative approaches is of the form $\mathbf{a}^T X Y^T \mathbf{b}$. This can be rewritten as :

$$\begin{aligned} \mathbf{a}^T X Y^T \mathbf{b} &= \sum_{k=1}^{n} \langle \mathbf{x}_k, \mathbf{a} \rangle \langle \mathbf{y}_k, \mathbf{b} \rangle \\ &= \sum_{k=1}^{n} \Big(\sum_{i=1}^{d_1} X_{i,k} a_i \cdot \sum_{j=1}^{d_2} Y_{j,k} b_j \Big) \\ &= \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} a_i b_j \Big(\sum_{k=1}^{n} X_{i,k} Y_{j,k} \Big) \\ &= \sum_{i,j=1}^{d_1,d_2} a_i b_j C_{ij}^{xy} \end{aligned} \quad (2)$$

Similarly, the constraints can also be rewritten as $\sum_{i,j=1}^{d_1} a_i a_j C_{ij}^{xx} = 1$ and $\sum_{i,j=1}^{d_2} b_i b_j C_{ij}^{yy} = 1$.

Maximizing this objective function, under the constraints, involves a careful selection of the vectors $\mathbf{a}$ and $\mathbf{b}$ such that $a_i b_j$ is high whenever C_{ij}^{xy} is high. So, every non-zero entry of the cross-covariance matrix restricts the choice of the projection directions. While this may not be a severe problem when the training data is clean, but this is very uncommon especially in the case of high dimensional data like text documents. Moreover, the inherent ambiguity of natural languages increases the chances of seeing a noisy word in any document. Every occurrence of a noisy word will have a non-zero contribution towards the covariance matrix making it dense, which in turn prevents the selection of appropriate projection directions.

In this section, we describe some techniques to recover the sparsity by removing the noisy entries from the covariance matrices. We break this task into two sub problems: computing an association score for every word pair and then using an appropriate strategy to identify the noisy pairs based on their weights. We explore multiple ways to address both the steps in the following two sections. For the sake of convenience and clarity, we describe our techniques in the context of cross-covariance matrix between English and Spanish language pair. But these techniques extend directly to monolingual covariance matrices, and to different language pairs as well.

3.1 Computing Word Pair Association

The first step in filtering out the noisy word co-occurrences is to use an appropriate measure to compute the strength of word pairs (English and Spanish words). This is a well studied problem and several association measures have been proposed in the NLP literature (Dunning, 1993; Inkpen and Hirst, 2002; Moore, 2004). These association measures can be divided into groups based on the statistics they use (Hoang et al., 2009). Here we explore a few of them for sparsifying the cross-covariance matrix.

3.1.1 Covariance

The first option is to use the cross-covariance matrix itself. As noted above, when the data matrix is centered, the cross-covariance of an English word (e_i) with a Spanish word (f_j) is given by $\sum_{k=1}^{n} X_{ik} Y_{jk}$. It measures the strength with which two words co-occur together. This measure uses information about the occurrence of a word pair in aligned documents and doesn't use other statistics such as 'how often this pair *doesn't* co-occur together' and so on.

3.1.2 Mutual Information

Association measures like covariance and Pointwise Mutual Information, which only use the frequency with which a word pair co-occurs, often overestimate the strength of low frequent words (Moore, 2004). On the other hand, measures like Log-likelihood ratio (Dunning, 1993) and Mutual Information (MI) use other statistics like the marginal probabilities of each of the words.

For any two words, e_i and f_j, let n_{11}, n_{10}, n_{01} and n_{00} denote the number of documents in which both the words co-occur, only English word occurs, only Spanish word occurs and none of the words occur. Then the Mutual Information of this word pair is given by:

$$\text{MI}(e_i, f_j) = \frac{1}{n} \sum_{i,j \in \{0,1\}} n_{ij} \log \frac{n_{ij} \times n}{n_i n_j} \quad (3)$$

where n_i and n_j denote the number of documents in which the English and the Spanish word occurs and n is the total number of documents. We treat the occurrence of a word in a document slightly different from others, we treat a word as occurring in a document if it has occurred more than its average frequency in the corpus. Log-likelihood ratio and the MI differ only in terms of the constant they use, so we use only MI in our experiments.

3.1.3 Yule's ω

Yule's ω is another popular association measure used in psychology (Reis and Judd, 2000). It uses same statistics used by Mutual Information but differs in the way in which they are combined. MI converts the frequencies into probabilities before computing the association measure where as Yule's ω uses the observed frequencies directly, and doesn't make any assumptions about the underlying probability distributions. Given the same interpretation of the variables as introduced in the previous section, the Yule's ω is estimated as:

$$\omega = \frac{\sqrt{n_{00} n_{11}} - \sqrt{n_{01} n_{10}}}{\sqrt{n_{00} n_{11}} + \sqrt{n_{01} n_{10}}} \quad (4)$$

This way of combining the frequencies bears similarity with the log-odds ratio.

3.1.4 Bilingual Dictionary

The above three association measures use the same training data that is available to compute the covariance matrices in CCA. Thus, their utility in bringing additional information, which is not captured by the covariance matrices, is arguable (our experiments show that they are indeed helpful). Moreover, they use document level co-occurrence information which is coarse compared to the co-occurrence at sentence level or the translational information provided by a bilingual dictionary. So, we use bilingual dictionaries as our final resource to weigh the word co-occurrences. Notice that, using bilingual information brings in information gleaned from an external corpus.

We use translation tables learnt using Giza++ (Och and Ney, 2003) on Europarl data set. Since the translation tables are asymmetric, we combine translation tables from both the directions. We first use a threshold on the conditional probability to filter out the low probability ones and then convert them into joint probabilities before combining. For each word pair (e_i, f_j), we compute the score as:

$$\frac{1}{2}\Big(P(e_i|f_j)P(f_j) + P(f_j|e_i)P(e_i)\Big)$$

While the first three association measures can also be applied to monolingual data, bilingual dictionary can't be used for weighting monolingual word pairs. So in this case, we use either of the above mentioned techniques for weighting monolingual word pairs.

3.2 Selection Strategies

The next step after computing association measure for all word pairs is to use them in selecting the pairs that need to be retained. In this section, we describe some approaches such as thresholding and matching for the word pair selection.

3.2.1 Thresholding

A straight forward way to remove the noisy word co-occurrences is to zero out the entries of the cross-covariance matrix that are lower than a threshold. To understand the motivation, consider the rewritten objective function of CCA, $\mathbf{a}^T XY^T \mathbf{b} = \sum_{ij} C_{ij}^{xy} a_i b_j$. This is linear in terms of the individual components of the cross-covariance matrix. So, if we want to remove some of the entries of the covariance matrix with minimal change in the value of the objective function, then the optimal choice is to sort the entries of the covariance matrix and filter out the less confident word pairs.

3.2.2 Relative Thresholding

While the thresholding strategy described in the above section is very simple, it is often biased by the frequent words. Since a frequent word co-occurs with other words often, it naturally tends to have high association with most of the other words. As a result, absolute thresholding tends to remove all the less frequent word pairs while leaving the co-occurrences of the frequent words untouched. Eventually, this may lead to zeroing out some of the rows or the columns of the cross-covariance matrix.

To circumvent this, we try thresholding at word level. For every English word, we choose a few Spanish words that have high association and vice versa. Since the nearest neighbour property is asymmetric, we take the union of all the selected word pairs. That is, we retain a word pair, if either the Spanish word is in the top ranked list of the English word or vice versa.

3.2.3 Maximal Matching

Though relative thresholding overcomes the problem of zeroing out entire rows or columns posed by direct thresholding, it is still biased by the frequent words. The high association measure of a frequent English word with many Spanish words, makes it a nearest neighbour for lot of Spanish words. One way to prevent this is to discourage an already selected English word from associating with a new Spanish word. This requires a global knowledge of all the selected pairs and can not be done by looking at the individual words, as is the case with the greedy strategy employed by the relative thresholding.

We use matching to solve this problem. We formulate the selection of the word pairs as a network flow problem (Jagarlamudi et al., 2011). The objective is to select word pairs that have high association measure while constraining each word to be associated with only a few words from other language. Let I_{ij} denote an indicator variable taking a value of

0 or 1 depending on if the word pair (e_i, f_j) is selected or not. We want each word to be associated with k words from other language, *i.e.*$\sum_j I_{ij} = k$ and $\sum_i I_{ij} = k$. Moreover, we want word pairs with high association score to be selected. We can encode this objective and the constraints as the following optimization problem:

$$\arg\max_{I} \sum_{i,j=1}^{d_1,d_2} C_{ij}^{xy} I_{ij} \quad (5)$$
$$\forall i \sum_j I_{ij} = k;\ \forall j \sum_i I_{ij} = k;\ \forall i,j\ I_{ij} \in \{0,1\}$$

If $k = 1$, then this problem reduces to a linear assignment problem and can be solved optimally using the Hungarian algorithm (Jonker and Volgenant, 1987). For other values of k, this can be solved by relaxing the constraint $I_{ij} \in \{0,1\}$ to $0 \leq I_{ij} \leq 1$. The optimal solution of the relaxed problem can be found efficiently using linear programming (Ravindra et al., 1993). The uni-modular nature of the constraints guarantees an integral solution (Schrijver, 2003), so relaxing the original integer problem doesn't introduce any error in the optimal solution.

3.2.4 Monolingual Augmentation

The above three selection strategies operate on the covariance matrices independently. In this section we propose to combine them. Specifically, we propose to augment the set of selected bilingual word pairs using the monolingual word pairs. We first use any of the above mentioned strategies to select bilingual and monolingual word pairs. Let I^{xy}, I^{xx} and I^{yy} be the binary matrices that indicate the selected word pairs based on the bilingual and monolingual association scores. Then the monolingual augmentation strategy updates I^{xy} in the following way:

$$I^{xy} \leftarrow \text{Binarize}(I^{xx} I^{xy} I^{yy})$$

i.e., we multiply I^{xy} with the monolingual selection matrices and then binarize the resulting matrix. Our monolingual augmentation is motivated by the following probabilistic interpretation:

$$P(x,y) = \sum_{x',y'} P(x|x')P(y|y')P(x',y')$$

which can be rewritten as $P \leftarrow T^x P (T^y)^T$ where T^x and T^y are monolingual state transition matrices.

3.3 Our Approach

In this section we summarize our approach for the task of finding aligned documents from a cross-lingual comparable corpora. The training phase involves finding projection directions for documents of both the languages. We compute the covariance matrices using the training data. Then we use any of the word association measures (Sec. 3.1) along with a selection criteria (Sec. 3.2) to recover the sparseness in either only the cross-covariance or all of the covariance matrices. Let I^{xy}, I^{xx} and I^{yy} be the binary matrices which represent the word pairs that are selected based on the chosen sparsification technique. Now, we replace the covariance matrices in Eq. 1 as follows: $C^{xx} \leftarrow C^{xx} \otimes I^{xx}$, $C^{yy} \leftarrow C^{yy} \otimes I^{yy}$ and $C^{xy} \leftarrow C^{xy} \otimes I^{xy}$ where $\otimes$ denotes the element-wise matrix product. Subsequently, we solve the generalized eigenvalue problem shown in Eq. 1 to obtain the projection directions. Let A and B be the matrices formed with top eigenvectors of Eq. 1 as the columns. These projection matrices are used to map documents into the interlingual representation. Such an interlingual representation is useful in many tasks like cross-lingual text categorization (Bel et al., 2003) multilingual web search (Gao et al., 2009) and so on.

During the testing, given an English document $\mathbf{x}$, finding an aligned Spanish document involves solving:

$$\arg\max_{\mathbf{y}} \frac{\mathbf{x}^T \Big((AB^T) \otimes I^{xy}\Big) \mathbf{y}}{\sqrt{\mathbf{x}^T \Big((AA^T) \otimes I^{xx}\Big) \mathbf{x}} \sqrt{\mathbf{y}^T \Big((BB^T) \otimes I^{yy}\Big) \mathbf{y}}}$$

If the documents are normalized before hand, then the above equation reduces to computing only the numerator.

4 Experiments

4.1 Experimental Setup

We experiment with the task of finding aligned documents from a cross-lingual comparable corpora. In this task, we are given comparable corpora consisting of two document collections, each in a different language. As the corpora are comparable, some documents in one collection have a comparable document in the other collection. The task is to recover

this hidden alignment. The recovered alignment is compared against the ground truth.

We evaluate our idea of sparsifying the covariance matrices incrementally. We first evaluate the effectiveness of our approach on synthetic data, as it enables us to systematically study the effect of noise. Subsequently, we evaluate each of the above discussed sparsification strategies on real world data sets. We have discussed four possible ways for computing word association measure and three approaches for word pair selection. That leaves us 12 different ways for sparsifying the covariance matrices, with each method having parameters to control the amount of sparseness. We use a small amount of development data for model selection and parameter tuning and choose a few promising models. Finally, we compare these selected models with state-of-the-art baselines on two language pairs and on two different data sets.

In each case, we use the training data to learn the projection directions. And then, for each of the test documents, we find the aligned document from other language. We report average accuracy of the top ranked document and also the Mean Reciprocal Rank (MRR) of the true aligned document.

4.2 Synthetic Data

We follow the generative story introduced in Bach and Jordan (2005) to generate synthetic multi-view data. Their method does not assume any correspondence between the feature dimensions of both the views. We modify their approach slightly so that we know the actual correspondence between the features. We use these true feature correspondences for sparsification of the cross-covariance matrix.

We first generate a d dimensional vector in the common latent space and then use the projection matrices to map it into the individual feature spaces as follows:

$$z \sim \mathcal{N}(0, I_d)$$
$$x|z \sim (W_1 z + \mu_1) + \eta\, \mathcal{N}(0, I_{d_1})$$
$$y|z \sim (W_1 z + \mu_2) + \eta\, \mathcal{N}(0, I_{d_2})$$

Notice that we use the same projection matrix W_1 for both the views, this ensures a one-to-one correspondence between the features of both the spaces. Moreover, we also introduce a parameter η which controls the amount of noise in the data.

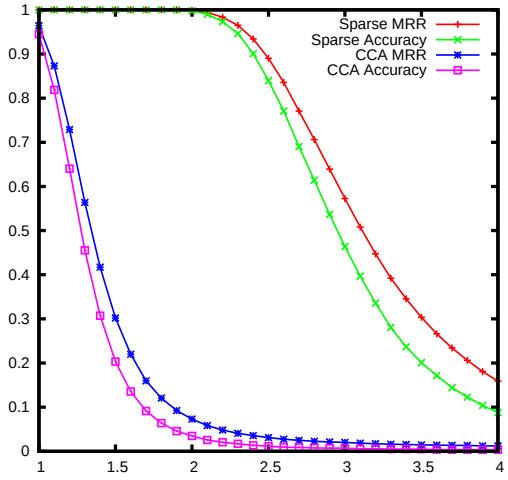


Figure 1: Accuracy of CCA and our sparsified version with the noise parameter.

We generate a total of 3000 pairs of points and use 2000 of them for training the models and the rest for evaluation. We use the true feature correspondences to form the cross-covariance selection matrix I^{xy} (Sec. 3.3). For this experiment, we use the full monolingual covariance matrices. We train both CCA and our sparse version on the training data and evaluate them on the test data. We repeat this multiple times and report the average accuracies. Fig. 1 shows the performance of CCA and our sparse CCA, as we vary the noise parameter η from 1 to 4. It is very clear that the sparse version performs significantly better than CCA. As the noise increases, the performance of CCA drops quickly. This experiment demonstrates a significant performance gain when the true correspondences are available. But this information is not available in the case of real world data sets, so we try to approximate it.

4.3 Model Selection

As we have discussed, there are several choices for computing the association measure and for selecting the word pairs to be retained. And each of them have sparsity parameters, giving raise to many possible models. For model selection, we use approximately 5000 document pairs collected from the Wikipedia between English and Spanish. We use the cross-language links provided as the ground truth. We tokenize the documents, retain only the most frequent 2000 words in each language and convert the docu-

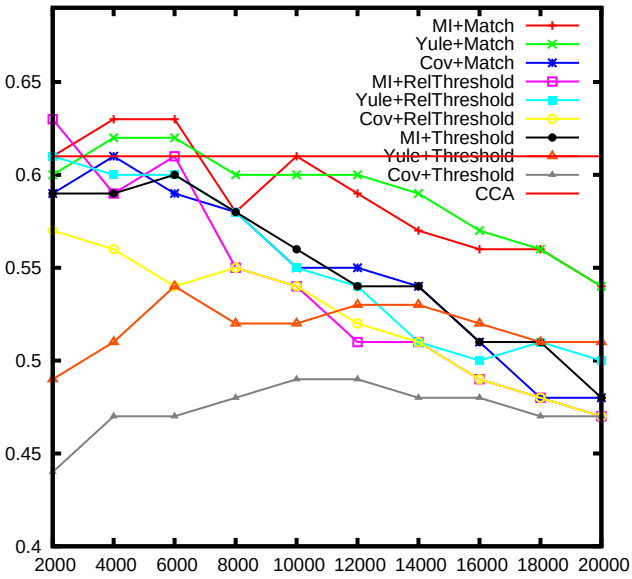


Figure 2: Comparison of the word association measures along with different selection criteria. The x-axis plots the number of non-zero entries in the covariance matrices and the y-axis plots the accuracy of top-ranked document.

ments into TFIDF vectors. We use 60% of the data for training different models and the rest for evaluating the models. We choose a few promising models based on this development set results and evaluate them on bigger data sets.

4.3.1 Selection Strategies

In the first experiment, we combine the three association measures, Covariance (Cov), MI and Yule's ω, with the three selection criteria, Threshold, Relative Threshold (RelThreshold) and Matching (Match). Fig. 2 shows the performance of these different combinations with varying levels of sparsity in the covariance matrices. The horizontal line represents the performance of CCA on this data set. We start with 2000 non-zero entries in the covariance matrices and experiment up to 20,000 non-zero entries. Since our data set has 2000 words in each language, 2000 non-zero entries in a covariance matrix implies that, on an average, every word is associated with only one word. This results in highly sparse covariance matrices.

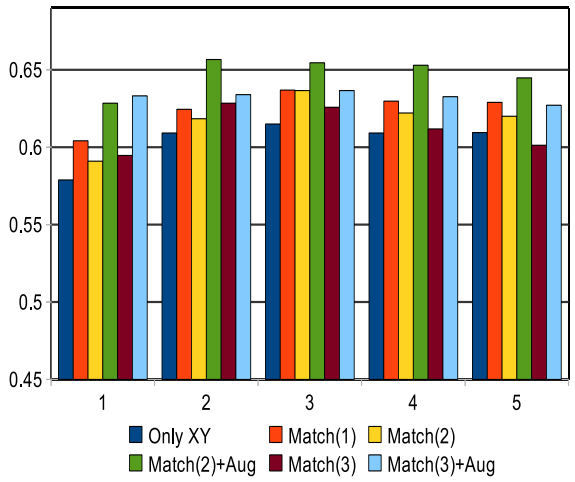
Overall, we observe that reducing the level of sparsity , *i.e.* selecting more number of elements in the covariance matrices, increases the performance slightly and then decreases again. From the figure, it seems that sparsifying the covariance matrices might help in improving the performance of the task. But it is interesting to note that not all the models perform better than CCA. In fact, both the models that achieve better scores use Matching as the selection criteria. This suggests that, apart from the weighting of the word pairs, appropriate selection of the word pairs is also equally important. In the rest of the experiments we mainly report results with Matching as the selection criterion. From this figure, we observe that Mutual Information and Yule's ω perform competitively but they consistently outperform models that use covariance as the association measure. So in the rest of the experiments we report results with MI or Yule's ω.

4.3.2 Amount of Sparsity

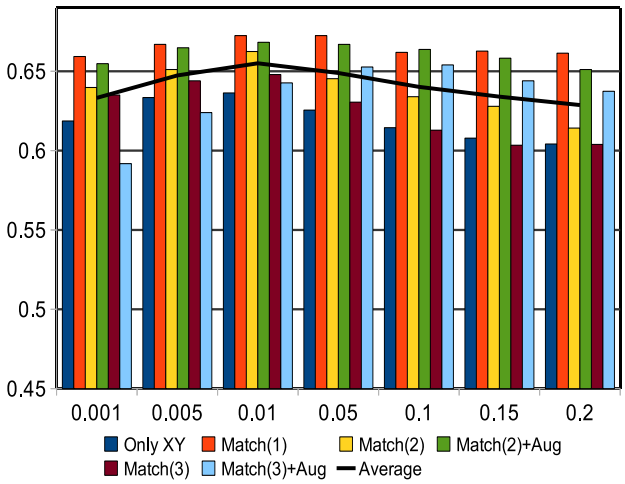
In the previous experiment, we used same level of sparsity for all the covariance matrices, *i.e.* same number of associations were selected for each word in all the three covariance matrices. In the following experiment, we use different levels of sparsity for the individual covariance matrices. Fig. 3 shows the performance of Yule+Match and Dictionary+Match combinations with different levels of sparsity. In the Yule+Match combination, we use Yule's ω association measure for weighting the word pairs and use matching for selection. In the Dictionary+Match combination, we use bilingual dictionary for sparsifying cross-covariance matrix, *i.e.* we keep all the word pairs whose conditional translation probability is above a threshold. And for monolingual word pairs, we use MI for weighting and matching for word pair selection.

For each level of sparsity of the cross-covariance matrix, we experiment with different levels of sparsity on the monolingual covariance matrices. 'Only XY' indicates we use the full monolingual covariance matrices. In 'Match(k)' runs, we allow each word to be associated with a total of k words (Eq. 5). 'Aug' indicates that we use monolingual augmentation to refine the sparsity of the cross-covariance matrix (Sec. 3.2.4).

From both the figures 3(a) and 3(b), we observe that 'Only XY' run (dark blue) performs poorly compared to the other runs, indicating that sparsifying all the covariance matrices is better than sparsifying only the cross-covariance matrix. In the



(a) Performance of Yule+Match combination. The x-axis plots the number of Spanish words selected per each English word and vice versa. This determines the sparsity of C^{xy}. Matching is used as selection criteria for all the covariance matrices.



(b) Performance of Dictionary+Match combination. The x-axis plots the threshold on bilingual translation probability and it determines the sparsity of C^{xy}. Matching is used to select *only* the monolingual sparsity.

Figure 3: Comparison of Yule+Match and Dictionary+Match combination with different levels of sparsity for the covariance matrices. In both the figures, the x-axis plots the sparsity of the cross-covariance matrix and for each value we try different levels of sparsity on the monolingual covariance matrices (which are grouped together). The description of these individual runs is provided in the relevant parts of the text. The y-axis plots the accuracy of the top-ranked document. CCA achieves 61% accuracy on this data set.

Yule+Match combination, Fig. 3(a), all the runs seem to be performing better when each English word is allowed to associate with 2 or 3 Spanish words and vice versa. Among different ways of selecting the monolingual word pairs, Match(2)+Aug performs better than the remaining runs. So we use Match(2)+Aug combination for the Yule's ω measure.

Unlike the Yule+Match combinations, there is no clear winner for Dictionary+Match combinations. First of all, the performance increase as we increase the translation probability threshold and then decreases again (indicated by the 'Average' performance in Fig. 3(b)). On an average, all the systems perform better with a threshold of 0.01, which we use in our final experiments. In this case, both Match(1) and Match(2)+Aug runs (orange and green bars respectively) perform competitively so we use both of these models in our final experiments.

In both the above experiments, the performance bars are very similar when we use MI instead of Yule and vice versa for weighting monolingual word pairs. Thus, to illustrate the main ideas we chose Yule's ω for the former combination and MI for the latter combination.

4.3.3 Promising Models

Based on the above experiments, we choose the following combinations for our final experiments. Yule(l)+Match(k), where $l \in \{2, 3\}$ is the number of Spanish words allowed for each English word and vice versa and k=2 is the number of monolingual word associations for each word. We also run both these combinations with monolingual augmentation, indicated by Yule(l)+Match(k)+Aug. For dictionary based weighting, Dictionary+Match(k), we choose a translation probability threshold of 0.01 and try $k \in \{1, 2\}$. Again, we run these combinations with monolingual augmentation identified by Dictionary+Match(k)+Aug.

4.4 Results

For our final results, we choose data in two language pairs (English-Spanish and English-German) from two different resources, Europarl (Koehn, 2005) and Wikipedia. For Europarl data sets, we artificially make them comparable by considering the first half

	Wikipedia				Europarl			
	English-Spanish		English-German		English-Spanish		English-German	
	Acc.	MRR	Acc.	MRR	Acc.	MRR	Acc.	MRR
CCA	0.776	0.852	0.570	0.699	0.872	0.920	0.748	0.831
OPCA	0.781	0.856	0.570	0.700	0.870	0.920	0.748	0.831
Yule(2)+Match(2)	0.798*	0.866*	0.576	0.703	0.901*	0.939*	0.780*	0.853*
Yule(2)+Match(2)+Aug	0.811*	0.876*	0.602*	0.723*	0.883	0.927	0.771*	0.847*
Yule(3)+Match(2)	0.803*	0.870*	0.572	0.700	0.856	0.907	0.747	0.830
Yule(3)+Match(2)+Aug	0.793*	0.861*	0.610*	0.726*	0.878$^+$	0.925$^+$	0.763$^+$	0.843*
Dictionary+Match(1)	0.811*	0.875*	**0.656***	**0.762***	**0.928***	**0.957***	**0.874***	**0.922***
Dictionary+Match(2)	0.811*	0.876*	0.623*	0.736*	0.923*	0.955*	0.853*	0.907*
Dictionary+Match(2)+Aug	**0.825***	**0.885***	0.630*	0.735*	0.897*	0.935*	0.866*	0.917*

Table 1: Performance of our models in comparison with CCA and OPCA on English-Spanish and English-German language pairs. * and $^+$ indicate statistical significance measured by paired t-test at p=0.01 and 0.05 levels respectively. When an improvement is significant at p=0.01 it is automatically significant at p=0.05 and hence is not shown.

of English document and the second half of its aligned foreign language document (Mimno et al., 2009). For Wikipedia data set, we use the cross-language link as the ground truth. For each of these data sets, we choose approximately 5000 aligned document pairs. We remove the stop words and keep all the words that occur in at least five documents. After the preprocessing, on an average, we are left with 4700 words in each language. Subsequently we convert the documents into their TFIDF representation.

In Platt *et al.* (2010), the authors compare different systems on the comparable document retrieval task and show that discriminative approaches work better compared to their generative counter parts. So, here we compare only with the state-of-the-art discriminative systems such as CCA and OPCA (Platt et al., 2010). For each of the systems, we report the average results of five-fold cross validation. We divide the data into 3:1:1 ratio for training, validation and test sets. The validation data set is used to select the best number of dimensions of the common sub space. For both CCA and our models, we set the regularization parameter λ to 0.3 which we found works well in a relevant but different experiments. For OPCA, we manually tried different regularization parameters ranging from 0.0001 to 1 and found that a value of 0.001 worked best.

The results are shown in Table 1. On these data sets, both CCA and OPCA performed competitively. OPCA takes advantage of the common vocabulary in both the languages. But in our data sets, vocabulary of both the languages is treated differently, so it is not surprising that they give almost the same results. From the results, it is clear that sparsifying the covariance matrices helps improving the accuracies significantly. In all the four data sets, the best performing method always used dictionary for cross-lingual sparsity selection. This indicates that using fine granular information such as a bilingual dictionary gleaned from an external source is very helpful in improving the accuracies. Among the models that rely solely on the training data, models that use monolingual augmentation performed better on Wikipedia data set, while models that do not use augmentation performed better on Europarl data sets. This suggests that, when the aligned documents are clean (closer to being parallel) the statistics computed from cross-lingual corpora are trustworthy. As the documents become comparable, we need to use monolingual statistics to refine the bilingual statistics. Moreover, these models achieve higher gains in the case of Wikipedia data set compared to the gains in Europarl. This conforms with our initial hunch that, when the training data is clean the covariance matrices tend to be less noisy.

5 Discussion

In this paper, we have proposed the idea of sparsifyng covariance matrices to improve bilingual pro-

jection directions. We are not aware of any NLP research that attempts to recover the sparseness of the covariance matrices to improve the projection directions. Our work is different from the sparse CCA (Hardoon and Shawe-Taylor, 2011; Rai and Daumé III, 2009) proposed in the Machine Learning literature. Their objective is to find projection directions such that the original documents are represented as a sparse vectors in the common sub-space. Another seemingly relevant but different direction is the sparse covariance matrix selection research (Banerjee et al., 2005). The objective in this work is to find matrices such that the *inverse* of the covariance matrix is sparse which has applications in Gaussian processes.

In this paper, we tried sparsification in the context of CCA only but our technique is general and can be applied to its variants like OPCA. Our experimental results show that using external information such as bilingual dictionaries which is gleaned from cleaner resources brings significant improvements. Moreover, we also observe that computing word pair association measures from the same training data along with an appropriate selection criteria can *also* yield significant improvements. This is certainly encouraging and in future we would like to explore more sophisticated techniques to recover the sparsity based on the training data itself.

6 Acknowledgments

We thank the anonymous reviewers for their helpful comments. This material is partially supported by the National Science Foundation under Grant No. 1139909.

References

Francis R. Bach and Michael I. Jordan. 2005. A probabilistic interpretation of canonical correlation analysis. Technical report, Dept Statist Univ California Berkeley CA Tech.

Lisa Ballesteros and W. Bruce Croft. 1996. Dictionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International Conference on Database and Expert Systems Applications*, DEXA '96, pages 791–801, London, UK. Springer-Verlag.

Onureena Banerjee, Alexandre d'Aspremont, and Laurent El Ghaoui. 2005. Sparse covariance selection via robust maximum likelihood estimation. *CoRR*, abs/cs/0506023.

Nuria Bel, Cornelis H. A. Koster, and Marta Villegas. 2003. Cross-lingual text categorization.

Hal Daume III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 407–412, Portland, Oregon, USA, June. Association for Computational Linguistics.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74, March.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 177–184, Morristown, NJ, USA. Association for Computational Linguistics.

Wei Gao, John Blitzer, Ming Zhou, and Kam-Fai Wong. 2009. Exploiting bilingual information to improve web search. In *Proceedings of Human Language Technologies: The 2009 Conference of the Association for Computational Linguistics*, ACL-IJCNLP '09, pages 1075–1083, Morristown, NJ, USA. ACL.

Aria Haghighi, Percy Liang, Taylor B. Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL-08: HLT*, pages 771–779, Columbus, Ohio, June. Association for Computational Linguistics.

David R. Hardoon and John Shawe-Taylor. 2011. Sparse canonical correlation analysis. *Journal of Machine Learning*, 83(3):331–353.

Ulf Hermjakob, Kevin Knight, and Hal Daumé III. 2008. Name translation in statistical machine translation - learning when to transliterate. In *Proceedings of ACL-08: HLT*, pages 389–397, Columbus, Ohio, June. Association for Computational Linguistics.

Hung Huu Hoang, Su Nam Kim, and Min-Yen Kan. 2009. A Re-examination of Lexical Association Measures. In *Proceedings of ACL-IJCNLP 2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, Singapore, August. Association for Computational Linguistics.

H. Hotelling. 1936. Relation between two sets of variables. *Biometrica*, 28:322–377.

Diana Zaiu Inkpen and Graeme Hirst. 2002. Acquiring collocations for lexical choice between near-synonyms. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition - Volume 9*, ULA '02, pages 67–76, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jagadeesh Jagarlamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned comparable corpora. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR*, volume 5993, pages 444–456, Milton Keynes, UK. Springer.

Jagadeesh Jagarlamudi, Hal Daume III, and Raghavendra Udupa. 2011. From bilingual dictionaries to interlingual document representations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 147–152, Portland, Oregon, USA, June. Association for Computational Linguistics.

R. Jonker and A. Volgenant. 1987. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 817–824, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 880–889, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robert C. Moore. 2004. On Log-Likelihood-Ratios and the Significance of Rare Events. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 333–340, Barcelona, Spain, July. Association for Computational Linguistics.

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

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

John C. Platt, Kristina Toutanova, and Wen-tau Yih. 2010. Translingual document representations from discriminative projections. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 251–261, Stroudsburg, PA, USA. Association for Computational Linguistics.

Piyush Rai and Hal Daumé III. 2009. Multi-label prediction via sparse infinite cca. In *Advances in Neural Information Processing Systems*, Vancouver, Canada.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 519–526, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2009. Learning phoneme mappings for transliteration without parallel data. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 37–45, Boulder, Colorado, June. Association for Computational Linguistics.

K. Ahuja Ravindra, L. Magnanti Thomas, and B. Orlin James. 1993. *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Inc.

Harry T Reis and Charles M Judd. 2000. *Handbook of Research Methods in Social and Personality Psychology*. Cambridge University Press.

Alexander Schrijver. 2003. *Combinatorial Optimization*. Springer.

Michael L. Littman Susan T. Dumais, Thomas K. Landauer. 1996. Automatic cross-linguistic information retrieval using latent semantic indexing. In *Working Notes of the Workshop on Cross-Linguistic Information Retrieval, SIGIR*, pages 16–23, Zurich, Switzerland. ACM.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res. (JAIR)*, 37:141–188.

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

Thuy Vu, AiTi Aw, and Min Zhang. 2009. Feature-based method for document alignment in comparable news corpora. In *EACL*, pages 843–851.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1128–1137, Uppsala, Sweden, July. Association for Computational Linguistics.