Pseudo-Aligned Multilingual Corpora

Fernando Diaz and Donald Metzler
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{fdiaz,metzler}@cs.umass.edu

Abstract

In machine translation, document alignment refers to finding correspondences between documents which are exact translations of each other. We define pseudo-alignment as the task of finding topical—as opposed to exact—correspondences between documents in different languages. We apply semisupervised methods to pseudo-align multilingual corpora. Specifically, we construct a topic-based graph for each language. Then, given exact correspondences between a subset of documents, we project the unaligned documents into a shared lower-dimensional space. We demonstrate that close documents in this lower-dimensional space tend to share the same topic. This has applications in machine translation and cross-lingual information analysis. Experimental results show that pseudo-alignment of multilingual corpora is feasible and that the document alignments produced are qualitatively sound. Our technique requires no linguistic knowledge of the corpus. On average when 10% of the corpus consists of exact correspondences, an on-topic correspondence occurs within the top 5 foreign neighbors in the lower-dimensional space while the exact correspondence occurs within the top 10 foreign neighbors in this this space. We also show how to substantially improve these results with a novel method for incorporating language-independent information.

1 Introduction

Electronic information is available in many different languages. If a user can only read Greek, then the amount of information available online is somewhat limited compared to a user who understands English. Therefore, in order to allow as many people to access as much information as possible, it is increasingly important to develop technologies that allow users to access information in a language-neutral fashion.

Two language technologies have been developed to tackle this task. First, *machine translation* systems attempt to bridge the language barrier by translating content on demand. This approach is appropriate when someone has a known-relevant document in hand. When this is not the case, *cross-lingual information retrieval* systems allow users to query a corpus in their native language and retrieve documents in a foreign language. The results can then either be manually or machine translated. We offer a hybrid approach which embeds all documents in multiple languages into a single semantic space. By providing a language-neutral embedding space, we can collectively analyze a foreign collection of documents without being constrained to document-based and query-based analysis. For example, a user may be interested in clustering or visualizing all documents in every language simultaneously.

We will now define our collection alignment problem. Assume that we are given two document collections. For example, consider one in English and one in Mandarin. In addition, we are given some training correspondences between documents we know are exact translations of each other. For example, assume we have a handful of English documents manually translated into Mandarin. Our task is, for each English or Mandarin document in the untranslated set, to find the topically most similar documents in the foreign corpus. This process results in a *pseudo-aligned corpus*.

Our approach aligns the underlying topical structures of two parallel collections.[1] Given a parallel corpus, the lexicon and distribution of terms within each side of the corpus will be quite different. However, since the corpus is parallel, the underlying topical structure is likely to be very similar regardless of the underlying language.

We conceptualize this topical structure in the form of a manifold over documents, where documents that are topically related are 'close' to each other on the manifold. Thus, we can view a corpus as a sample from some underlying manifold. We are interested in the case where the topical distributions between languages are very similar. Here, our working hypothesis is that the true underlying topical manifolds of any two languages are isomorphic.

We use techniques from spectral graph theory to automatically pseudo-align documents in different languages. Unlike machine translation systems, which focus on exact 1-to-1

[1] A *multilingual corpus* is a set of documents C such that each document is written in one of a set of languages $L = \{l_1, l_2, ...\}$. A *parallel corpus* has the additional property that for every document $d_i \in C$, there are $|L| - 1$ other documents in C that are exact translations of d_i into the other languages in L.

alignments of documents or sentences, we instead focus on a looser sense of alignment, based on topical relevance. Our results show that it is possible to recover topic and exact alignments of documents using a reasonably small set of training examples and very naïve linguistic processing. We also show how to improve these results with a novel method for incorporating language-independent information.

2 Related Work

Parallel corpora are a fundamental concept in machine translation. Traditionally, the alignment problem focuses on aligning sentences between two documents known to be exact translations [Gale and Church, 1993]. Statistical machine translation systems require this level of granularity to learn relationships between words in different languages. Our approach relaxes both the granularity and exact-translation constraints.

Oftentimes, we know parallel corpora exist but do not have the correspondences. This happens frequently on the world wide web where entire hierarchies may be represented in several languages. The solution to this problem usually requires inspecting and aligning URLs and structural tags in the documents [Resnik and Smith, 2003]. While this approach works well for structured and explicitly-linked data, when this information is missing or inexact, the solution may not work. Our approach only requires relationships between the content within a language and is robust to noise.

Another alternative to re-alignment uses external dictionaries to create probabilistic relationships between unaligned documents [Resnik and Smith, 2003]. While this technique is applicable to our task, we are interested in methods which do not require external resources such as dictionaries.

Our work is also related to the task of aligning multidimensional data sets. When viewing documents as, say, English term vectors and Mandarin term vectors, we can use techniques such as canonical correlation analysis or Gaussian processes to compute a transformation between the spaces [Hardoon *et al.*, 2004; Shon *et al.*, 2006]. Correspondences and translations can also be addressed in terms of graphical models [Barnard *et al.*, 2003]. Solutions using spectral graph theory are the most related to our work [Carcassoni and Hancock, 2003; Ham *et al.*, 2005; Shon *et al.*, 2006; Verbeek and Vlassis, 2006]. We apply these spectral techniques and extend them to include manifold-independent information.

3 Collection Alignment

Our procedure for aligning corpora consists of two phases: representing monolingual document collections and aligning the monolingual representations. In the first phase, we consider a graph-based representation of the document collection. Graphs provide intuitive and flexible collection models suitable for a variety of tasks such as classification and retrieval [Diaz, 2005; Zhu *et al.*, 2003]. The second phase is to find topically similar nodes in the foreign graph using labeled document alignments. We employ spectral graph theory to project documents in all languages into a single embedding space and align documents using distances in this joint embedding space.

3.1 Representing Document Collections

Graph-based representations of document collections view documents as nodes in a graph. Edges in this graph exist between documents which share a property such as topic, genre, author, etc. Because we are interested in topical alignment of collections, we will be focusing on topical edges. In this section, we will be discussing one method of detecting topical relationships. Although others certainly exist, graph-based representations have consistent behavior across affinity measures [Diaz, 2005].

Given a corpus containing n documents and $|V|$ terms, one of the most popular document representations is the length-$|V|$ term vector. Constructing the vector often requires a term-weighting scheme such as tf.idf. In our work, we will assume that document vectors are language models (multinomial term distributions) estimated using the document text [Croft and Lafferty, 2003]. By treating documents as probability distributions, we can use distributional affinity to detect topical relatedness between documents. Specifically, we use the multinomial diffusion kernel [Lafferty and Lebanon, 2005]. Given two documents i and j, the affinity is measured between the two distributions, θ_i and θ_j, as

$$\mathcal{K}(\theta_i, \theta_j) \quad = \quad \exp\left(-t^{-1} \arccos^2\left(\sqrt{\theta_i} \cdot \sqrt{\theta_j}\right)\right) \quad (1)$$

where t is a parameter controlling the decay of the affinity. The diffusion kernel has been shown to be a good affinity metric for tasks such as text classification and retrieval.

A document graph for a particular language, then, is constructed by treating the n documents as nodes and, for each document, adding undirected, weighted edges to the k nearest neighbors as measured by the diffusion kernel. We represent these a document graph as the $n \times n$ adjacency matrix W. In our experiments, we fix $t = 0.50$ and $k = 25$. We use a simple maximum likelihood estimate for the document language models.

3.2 Functions on Graphs

Because our alignment algorithm uses results from spectral clustering, we will briefly review some fundamentals before presenting our solution. A more thorough treatment of the material can be found in other sources [Chung, 1997].

We define a function f over the nodes of a graph as a length-n vector. We can measure the smoothness of this function as $\sum_{ij} W_{ij}(f_i - f_j)^2$. This is known as the Dirichlet sum and computes the difference in the function value between connected nodes.

The Dirichlet sum can be written as $f^{\mathsf{T}}(D - W)f$ where D is a diagonal matrix such that $D_{ii} = \sum_j W_{ij}$. The matrix $\Delta = D - W$ is known as the combinatorial Laplacian. We can introduce alternative Laplacians to provide different measures of smoothness. In this paper, we will always use the approximate Laplace-Beltrami operator [Lafon, 2004]. This is defined as,

$$\Delta \quad = \quad I - \hat{D}^{-1/2}\hat{W}\hat{D}^{-1/2} \quad (2)$$

where we use the normalizing affinity matrix $\hat{W} = D^{-1}WD^{-1}$ with $\hat{D} = \sum_{j=1}^{n} \hat{W}_{ij}$. This approximation provides a density normalization effect that we have found important when dealing with document collections.

The k eigenvectors associated with the lowest non-zero eigenvalues of the Laplacian represent the functions f minimizing the Dirichlet sum. In turn these eigenvectors can be used to embed documents in a lower dimensional space [Belkin and Niyogi, 2002]. If we let E represent the $n \times k$ matrix of these eigenvectors, we can represent each document i using the corresponding row vector of E. We then compute the Euclidean distance between documents in the k-dimensional space,

$$d^2(x_i, x_j) = \sum_k (E_{ik} - E_{jk})^2 \quad (3)$$

3.3 Aligning Collections

We now define our collection alignment problem. Assume that we are given two document graphs represented by the $n \times n$ adjacency matrices W^x and W^y. In addition, we are given $m < n$ training correspondences between documents we know are exact translations of each other. We can reorganize the adjacency matrices so that the indexes of corresponding documents match and are located in the $m \times m$ upper left blocks, W^x_{ll} and W^y_{ll}. Our task is to find the most topically similar documents for the unlabeled $2(n-m)$ documents.

We use the manifold alignment method proposed by Ham *et al* [Ham *et al.*, 2005]. Specifically, we are interested in finding the functions f and g minimizing the following objective,

$$C(f, g) = \frac{f^\mathsf{T}\Delta^x f + g^\mathsf{T}\Delta^y g}{f^\mathsf{T} f + g^\mathsf{T} g} \quad (4)$$

such that $f_i = g_i$ for $i < m$. The pairs of functions minimizing this objective can be used to project documents into a single lower-dimensional space.

Although both Laplacians, Δ^x and Δ^y, are the same size, the indexes $m \leq i < n$ refer to potentially different documents. Therefore, we build adjacency matrices with three sets of vertices: the first set of vertices is common between languages; these are the training instances with known alignment ($0 \leq i < m$). The second set is documents from language x with unknown alignment ($m \leq i < n$), and the third set is language y with unknown alignment ($n \leq i < 2n-m$). This results in the $(2n-m) \times (2n-m)$ matrices,

$$\hat{\Delta}^x = \begin{bmatrix} \Delta^x_{ll} & \Delta^x_{lu} & 0 \\ \Delta^x_{ul} & \Delta^x_{uu} & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (5)$$

$$\hat{\Delta}^y = \begin{bmatrix} \Delta^y_{ll} & 0 & \Delta^y_{lu} \\ 0 & 0 & 0 \\ \Delta^y_{ul} & 0 & \Delta^y_{uu} \end{bmatrix} \quad (6)$$

Notice that we are augmenting these graphs so that there are no edges to new nodes. We will see in Section 3.4 how to incorporate language-independent knowledge we might have about the relationship to these foreign documents.

We can rewrite Equation 4 using these augmented matrices,

$$C(h) = \frac{h^\mathsf{T}\Delta^z h}{h^\mathsf{T} h} \quad (7)$$

where $h = [f^\mathsf{T} g^\mathsf{T}]^\mathsf{T}$ and we define the composite Laplacian matrix, $\Delta^z = \hat{\Delta}^x + \hat{\Delta}^y$. This can be seen as using the combined Laplacian, Δ^z, of a new graph with $2n-m$ nodes.

When viewing alignment as analyzing a larger graph, we notice that Δ^z contains zero submatrices between unaligned nodes across languages. This is problematic since graph Laplacian techniques exploit link structure to detect topics. In order to address this issue, we "seed" these submatrices with predicted alignments from a simple baseline. We represent the $2(n-m)$ unaligned documents in an m-dimensional space. The elements of each document vector represent the affinity with the training documents. That is, we use the $(n-m) \times m$ lower left submatrices of W^x and W^y. We calculate the seed affinities by L_2 normalizing the rows and computing $W^x_{ul}(W^y_{ul})^\mathsf{T}$. This $(n-m) \times (n-m)$ matrix defines our initial predictions of the alignments between unaligned documents. We will refer to this as our baseline in experiments. In the case of these experiments, we place the prediction matrix in the middle-right/lower-middle blocks of $\hat{W}^x$ and $\hat{W}^y$.

We can align documents by first projecting all $2n-m$ documents into a lower-dimensional space and then computing distances in that lower-dimensional space. With enough labeled instances, the projection should improve the baseline predictions. We use the Laplacian-based projection method described in Section 3.2. Given a document x_i in language x, its predicted aligned pair in language y is the closest document in the embedding space. Highly ranked documents, then, are likely to be topically related.

3.4 Incorporating Language-Independent Information

In many cases, documents contain language-independent information which can be exploited for alignment. Examples include named entities, hyperlinks, and time-stamps. In this section, we extend Ham's alignment algorithm to consider such manifold-independent information. Specifically, we exploit the temporal information present in the document.

Recall that we viewed our alignment as using the combined Laplacian, Δ^z, of a new graph with $2n-m$ nodes. We would like to consider a second graph over these $2n-m$ nodes incorporating the language-independent knowledge. This graph will be defined so that edges occur when two documents share the same date; call this unweighted adjacency matrix W^t. This gives us two Laplacians, Δ^z and Δ^t over the large graph. We then measure the smoothness of the function h on both graphs,

$$C^t(h) = \frac{\lambda h^\mathsf{T}\Delta^z h + (1-\lambda) h^\mathsf{T}\Delta^t h}{h^\mathsf{T} h} \quad (8)$$

where the parameter λ allows us to weight the temporal information. Here, our solution falls from embedding documents in a lower dimensional space defined by the lowest non-constant eigenvectors of $\lambda\Delta^z + (1-\lambda)\Delta^t$.

1. compute $n \times n$ affinity matrices for languages x and y
2. add the 25 nearest neighbors for each document to W^x and W^y
3. compute the Laplacians, Δ^x and Δ^y
4. compute the predicted alignments
5. construct the combined Laplacian, Δ^z
6. if language-independent information exists interpolate $(1-\lambda)\Delta^z + \lambda\Delta^t$
7. compute the k eigenvectors associated with the smallest non-zero eigenvalues; stack in matrix E

n	number of documents in one side of the parallel corpus
k	dimensionality of the joint embedding space
λ	interpolation parameter for language-independent information
E	$n \times k$ projection of all documents into k-dimensional space

Figure 1: Pseudo-alignment algorithm. Input are k and λ. The output is a set of distance between all unlabeled documents. The closest pairs represent predicted alignments.

When we evaluate our algorithms using parallel corpora, this temporal information is powerful but unrealistic. Documents with shared topics will rarely have exactly the same date. Therefore, we consider a corruption of the date information in our corpus. We accomplish this by corrupting the date information through the following process: for each document i, select a date d from a Gaussian distribution whose mean is the date of document and a variance, σ. Select a document j uniformly from amongst all of the documents on that date. Construct an edge between i and j. Repeat this process 50 times for each document. The parameter σ allows us to control the error in establishing links between documents. A low σ will result in constructing edges to 50 nodes which share the same date as i; a high σ will result in constructing edges to 50 nodes less temporally local to i. This has the effect of modeling documents on the same topic as being published on different but close dates.

We present a summary of our alignment algorithm in Figure 1.

4 Methods and Materials

4.1 Corpora

Parallel corpora allow us to evaluate the document-level alignment for collections where the topical distributions are identical. We used two parallel corpora: an Arabic-English corpus of United Nations documents and an English-Mandarin corpus of newswire documents. The Arabic-English corpus consists of 30K United Nations documents manually translated into both languages [Ma *et al.*, 2004]. Because some dates were under-represented or missing, we only used documents between 1994 and 1999. The English-Mandarin corpus consists of 50K Chinese newswire documents published between August and September 2003 and their machine translated representations in English [Fiscus and Wheatley, 2004].

The English and Arabic sides of the corpora were tokenized on whitespace and punctuation. No stopping or stemming was performed. The Chinese corpora was tokenized using character unigrams. No additional segmentation or analysis was performed. After tokenization, documents were indexed using the Indri retrieval system [Strohman *et al.*, 2004]. We use only date stamps (not time stamps) as our language-independent information.

One concern we had when using parallel corpora was that the graph structures would be identical. We found that, even for our machine translated corpus, the graphs were quite different. Nevertheless, we conducted a set of experiments where random subsets of the nodes were removed from each side of the corpus. This is equivalent to having non-parallel collections with identical topical distributions.

4.2 Evaluation

We train our algorithms by providing m example correspondences randomly selected from the collection; in our experiments, we present the number of training correspondence as a fraction of the corpus. We evaluate the re-alignment of parallel corpora using two measures. First, we consider the mean reciprocal rank (MRR) of the true match. That is, we compute distance from a document in language x to all documents in language y; the reciprocal rank of the true translation of this document gives us the score for this document. We use the mean reciprocal rank over all $2(n-m)$ testing documents.

We noticed that even at few training correspondences, though the MRR was quite low (on average the true translation in the top 100 documents), the qualitative matches appeared quite good. For example, the closest neighbors to a document about Sri Lanka—while not including the exact translation—contained documents about Sri Lanka. Because our qualitative analysis suggested that MRR was underrepresenting our performance, we wanted to evaluate the topical alignment. Fortunately, a subset of the English-Mandarin corpus contains assessments for topical equivalence between documents. We therefore adapted the MRR measure to look for the top ranking on-topic document; we refer to this as TMRR.

5 Results

Our first set of experiments investigates the performance of our algorithms with respect to the training alignments. The number of eigenvectors was fixed at 300. Figure 2 depicts learning evaluated by MRR for the Arabic-English and English-Mandarin corpora and TMRR for the English-Mandarin corpus. The baseline algorithm uses only the distances to the training documents to predict alignments. Our alignment algorithm uses both these predictions as well as information about the relationship between unaligned documents.

The results in Figure 2 demonstrates that our alignment algorithm improves the baseline at few training documents. However, as the training size increases, this improvement disappears. We speculate that this task is such that, after a certain point, the number of training alignments provide enough information to adequately distinguish unaligned documents; the additional information encoded in the matrices W^x_{uu} and

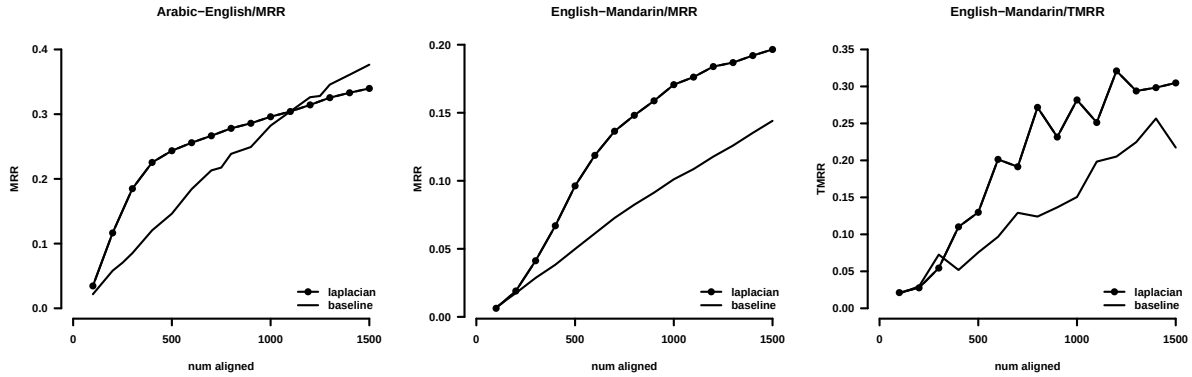

Figure 2: Mean reciprocal rank of the true translation for Arabic-English alignment (left) and English-Mandarin alignment (center). Topic mean reciprocal rank for English-Mandarin alignment (right). All algorithms used 300 eigenvectors.

W^y_{uu} do not add any discriminating information. In fact, because the Laplacian-based alignment technique discards information in the projection, performance may suffer. This can be seen in performance curves for the Arabic-English corpus. Nevertheless, when training data is sparse, the structure extracted by the Laplacian-based technique can be leveraged to improve on the baseline.

In all cases, both document-level and topical alignment are feasible even when only using content information. For example, at 500 training examples, we get the true alignment in the top 4 for the Arabic-English corpus and the top 10 for the English-Mandarin corpus. When looking at topical alignment, we can get an on-topic document in the 7 for the same number of training instances.

Our first experiments evaluated the realignment of parallel corpora. We were interested in testing the robustness of our techniques to non-parallel corpora. In order to accomplish this, we first fixed the number of training correspondences, m, at 1000. We also fixed the number of testing correspondences at 1000. We then added 20000 documents from each collection. These documents were selected such that some fraction of them were included as pairs of aligned documents. The remaining fraction were randomly sampled, potentially unaligned documents from the collections. Varying the fraction of unaligned documents in the 20000, we plotted the MRR for the test correspondences in Figure 3. We notice that both our baseline and new algorithm are not effected by the addition of unaligned documents.

We evaluated our temporal alignment using several values for σ (the date corruption parameter); we present results for the values $\{1, 2, 5\}$. We were interested in the performance over various values of λ. Fixing the training correspondences at 10000 and number of eigenvalues at 150. we varied the value of λ. We present these results in Figure 4. Here, the improvements gained by introducing temporal information are very dependent on the value of λ. Obviously, at low values, the algorithms will reduce to the text-based alignment. However, the reduction in performance observed at the higher values for λ are likely due to documents being ranked exclusively by their temporal proximity.

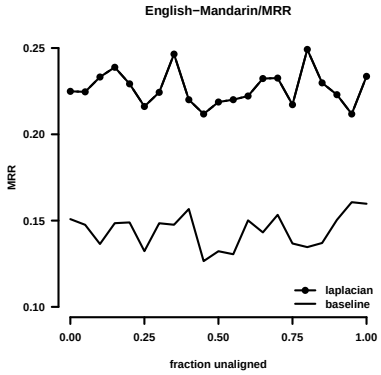

Figure 3: Performance as a function of unaligned documents added to the collection. We fixed the training and testing set sizes to 1000 and added 20000 documents from each side of the corpus. Of these 20000, some fraction were not required to be aligned pairs.

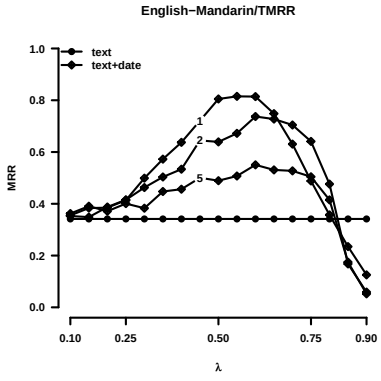

Figure 4: Incorporation of language-independent information. We fixed the number of training correspondences to be .20 of the collection and evaluated performance as a function of the weight, λ, placed on the language-independent information.

We caution that the temporal corruption process introduces temporal dimensions to topics which are potentially atemporal. For example, there is no reason to believe that two documents (one in English and one in Mandarin) about cooking will be published on or around the same date. With this in mind, our experiment provide suggestive results with respect to the merit of temporal information for topical alignment. Certainly in the cases where the user is interested in temporally salient topics, date information will be invaluable.

6 Conclusion

In this paper, we hypothesized that the true underlying topical manifold of different languages are isomorphic. We treated parallel corpora as samples from this underlying manifold and represented the manifold structure using graphs.

We then described a semi-supervised algorithm for aligning parallel corpora. Given a set of correspondences, we apply a spectral graph technique to embed the documents in a lower dimensional space. This essentially clusters the documents in each language and uses the correspondences to define a joint embedding over the entire space. Using this joint embedding, we then evaluated our alignment using a document-matching measure (MRR) and as well as a topic-matching measure (TMRR). We demonstrated that our technique can retrieve the true translation of a document at relatively high ranking and an on-topic document near rank 1 at low numbers of training alignments.

This work suggests several interesting directions. First, although we considered only two languages, our framework easily allows the incorporation of additional languages. This would allow one to leverage topic information from different languages when defining the lower-dimensional topic space. Second, we adopted parallel corpora for evaluation reasons. The alignment algorithm does not require that the unlabeled data be parallel. In fact, it would be very helpful to explore the robustness of our alignment method when the two collection have very different topic distributions. Would they techniques perform well if documents in languages x and y were drawn from arbitrary, non-parallel collections? Finally, these techniques can also be applied to multimedia situations where we want to align documents in different media.

7 Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #CNS-0454018 , and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

[Barnard *et al.*, 2003] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003.

[Belkin and Niyogi, 2002] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

[Carcassoni and Hancock, 2003] Marco Carcassoni and Edwin R. Hancock. Spectral correspondence for point pattern matching. *Pattern Recognition*, 36(1):193–204, 2003.

[Chung, 1997] Fan R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[Croft and Lafferty, 2003] W. Bruce Croft and John Lafferty. *Language Modeling for Information Retrieval*. Kluwer Academic Publishing, 2003.

[Diaz, 2005] Fernando Diaz. Regularizing ad hoc retrieval scores. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 672–679, New York, NY, USA, 2005. ACM Press.

[Fiscus and Wheatley, 2004] J. Fiscus and B. Wheatley. Overview of the tdt 2004 evaluation and results. In *TDT5*, 2004.

[Gale and Church, 1993] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19(1):75–102, 1993.

[Ham *et al.*, 2005] Jihun Ham, Daniel Lee, and Lawrence Saul. Semisupervised alignment of manifolds. In Robert G. Cowell and Zoubin Ghahramani, editors, *Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, pages 120–127. Society for Artificial Intelligence and Statistics, 2005.

[Hardoon *et al.*, 2004] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, December 2004.

[Lafferty and Lebanon, 2005] John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *J. Mach. Learn. Res.*, 6:129–163, 2005.

[Lafon, 2004] Stephane Lafon. *Diffusion Maps and Geometric Harmonics*. PhD thesis, Yale University, 2004.

[Ma *et al.*, 2004] Xiaoyi Ma, Jinxi Xu, Alexander Fraser, John Makhoul, Mohamed Noamany, and Ghada Osman. Un arabic english parallel text version 1.0 beta. LDC2004E13, 2004.

[Resnik and Smith, 2003] Philip Resnik and Noah A. Smith. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380, 2003.

[Shon *et al.*, 2006] Aaron Shon, Keith Grochow, Aaron Hertzmann, and Rajesh Rao. Learning shared latent structure for image synthesis and robotic imitation. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA, 2006.

[Strohman *et al.*, 2004] Trevor Strohman, Donald Metzler, Howard Turtle, and W. B. Croft. Indri: A language model-based serach engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*, 2004.

[Verbeek and Vlassis, 2006] Jakob J. Verbeek and Nikos Vlassis. Gaussian fields for semi-supervised regression and correspondence learning. *To appear in Pattern Recognition*, 2006.

[Zhu *et al.*, 2003] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *The Twentieth International Conference on Machine Learning (ICML-2003)*, 2003.