From Bilingual Dictionaries to Interlingual Document Representations

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

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

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

Abstract

Mapping documents into an interlingual representation can help bridge the language barrier of a cross-lingual corpus. Previous approaches use aligned documents as training data to learn an interlingual representation, making them sensitive to the domain of the training data. In this paper, we learn an interlingual representation in an unsupervised manner using only a bilingual dictionary. We first use the bilingual dictionary to find candidate document alignments and then use them to find an interlingual representation. Since the candidate alignments are noisy, we develop a robust learning algorithm to learn the interlingual representation. We show that bilingual dictionaries generalize to different domains better: our approach gives better performance than either a word by word translation method or Canonical Correlation Analysis (CCA) trained on a different domain.

1 Introduction

The growth of text corpora in different languages poses an inherent problem of aligning documents across languages. Obtaining an explicit alignment, or a different way of bridging the language barrier, is an important step in many natural language processing (NLP) applications such as: document retrieval (Gale and Church, 1991; Rapp, 1999; Ballesteros and Croft, 1996; Munteanu and Marcu, 2005; Vu et al., 2009), Transliteration Mining (Klementiev and Roth, 2006; Hermjakob et al., 2008; Udupa et al., 2009; Ravi and Knight, 2009) and Multilingual Web Search (Gao et al., 2008; Gao et al., 2009). Aligning documents from different languages arises in all the above mentioned problems. In this paper, we address this problem by mapping documents into a common subspace (interlingual representation)[1]. This common subspace generalizes the notion of vector space model for cross-lingual applications (Turney and Pantel, 2010).

There are two major approaches for solving the document alignment problem, depending on the available resources. The first approach, which is widely used in the Cross-lingual Information Retrieval (CLIR) literature, uses bilingual dictionaries to translate documents from one language (source) into another (target) language (Ballesteros and Croft, 1996; Pirkola et al., 2001). Then standard measures such as cosine similarity are used to identify target language documents that are close to the translated document. The second approach is to use training data of aligned document pairs to find a common subspace such that the aligned document pairs are maximally correlated (Susan T. Dumais, 1996; Vinokourov et al., 2003; Mimno et al., 2009; Platt et al., 2010; Haghighi et al., 2008) .

Both kinds of approaches have their own strengths and weaknesses. Dictionary based approaches treat source documents independently, *i.e.*, each source language document is translated independently of other documents. Moreover, after translation, the relationship of a given source document with the rest of the source documents is ignored. On the other hand, supervised approaches use all the source and target language documents to infer an interlingual

[1]We use the phrases "common subspace" and "interlingual representation" interchangeably.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 147–152,
Portland, Oregon, June 19-24, 2011.

representation, but their strong dependency on the training data prevents them from generalizing well to test documents from a different domain.

In this paper, we propose a technique that combines the advantages of both these approaches. At a broad level, our approach uses bilingual dictionaries to identify initial noisy document alignments (Sec. 2.1) and then uses these noisy alignments as training data to learn a common subspace. Since the alignments are noisy, we need a learning algorithm that is robust to the errors in the training data. It is known that techniques like CCA overfit the training data (Rai and Daumé III, 2009). So, we start with an unsupervised approach such as Kernelized Sorting (Quadrianto et al., 2009) and develop a supervised variant of it (Sec. 2.2). Our supervised variant learns to modify the within language document similarities according to the given alignments. Since the original algorithm is unsupervised, we hope that its supervised variant is tolerant to errors in the candidate alignments. The primary advantage of our method is that, it does not use any training data and thus generalizes to test documents from different domains. And unlike the dictionary based approaches, we use all the documents in computing the common subspace and thus achieve better accuracies compared to the approaches which translate documents in isolation.

There are two main contributions of this work. First, we propose a discriminative technique to learn an interlingual representation using *only* a bilingual dictionary. Second, we develop a supervised variant of Kernelized Sorting algorithm (Quadrianto et al., 2009) which learns to modify within language document similarities according to a given alignment.

2 Approach

Given a cross-lingual corpus, with an underlying unknown document alignment, we propose a technique to recover the hidden alignment. This is achieved by mapping documents into an interlingual representation. Our approach involves two stages. In the first stage, we use a bilingual dictionary to find initial candidate noisy document alignments. The second stage uses a robust learning algorithm to learn a common subspace from the noisy alignments identified in the first step. Subsequently, we project all the documents into the common subspace and use maximal matching to recover the hidden alignment. During this stage, we also learn mappings from the document spaces onto the common subspace. These mappings can be used to convert any new document into the interlingual representation. We describe each of these two steps in detail in the following two sub sections (Sec. 2.1 and Sec. 2.2).

2.1 Noisy Document Alignments

Translating documents from one language into another language and finding the nearest neighbours gives potential alignments. Unfortunately, the resulting alignments may differ depending on the direction of the translation owing to the asymmetry of bilingual dictionaries and the nearest neighbour property. In order to overcome this asymmetry, we first turn the documents in both languages into bag of translation pairs representation.

We follow the feature representation used in Jagarlamudi and Daumé III (2010) and Boyd-Graber and Blei (2009). Each translation pair of the bilingual dictionary (also referred as a dictionary entry) is treated as a new feature. Given a document, every word is replaced with the set of bilingual dictionary entries that it participates in. If D represents the TFIDF weighted term $\times$ document matrix and T is a binary matrix matrix of size *no_of_dictionary_entries* $\times$ *vocab_size*, then converting documents into a bag of dictionary entries is given by the linear operation $X^{(t)} \leftarrow TD$.[2]

After converting the documents into bag of dictionary entries representation, we form a bipartite graph with the documents of each language as a separate set of nodes. The edge weight W_{ij} between a pair of documents $x_i^{(t)}$ and $y_j^{(t)}$ (in source and target language respectively) is computed as the Euclidean distance between those documents in the dictionary space. Let π_{ij} indicate the likeliness of a source document $x_i^{(t)}$ is aligned to a target document $y_j^{(t)}$. We want each document to align to at least one document from other language. Moreover, we want to encourage similar documents to align to each other. We can formulate this objective and the constraints as the following minimum cost flow

[2]Superscript (t) indicates that the data is in the form of bag of dictionary entries

problem (Ravindra et al., 1993):

$$\arg\min_{\pi} \sum_{i,j=1}^{m,n} W_{ij}\pi_{ij} \quad (1)$$
$$\forall i \sum_j \pi_{ij} = 1 \; ; \; \forall j \sum_i \pi_{ij} = 1$$
$$\forall i,j \quad 0 \leq \pi_{ij} \leq C$$

where C is some user chosen constant, m and n are the number of documents in source and target languages respectively. Without the last constraint ($\pi_{ij} \leq C$) this optimization problem always gives an integral solution and reduces to a maximum matching problem (Jonker and Volgenant, 1987). Since this solution may not be accurate, we allow many-to-many mapping by setting the constant C to a value less than one. In our experiments (Sec. 3), we found that setting C to a value less than 1 gave better performance analogous to the better performance of soft Expectation Maximization (EM) compared to hard-EM. The optimal solution of Eq. 1 can be found efficiently using linear programming (Ravindra et al., 1993).

2.2 Supervised Kernelized Sorting

Kernelized Sorting is an unsupervised technique to align objects of different types, such as English and Spanish documents (Quadrianto et al., 2009; Jagaralmudi et al., 2010). The main advantage of this method is that it *only* uses the **intra**-language document similarities to identify the alignments across languages. In this section, we describe a supervised variant of Kernelized Sorting which takes a set of candidate alignments and learns to modify the intra-language document similarities to respect the given alignment. Since Kernelized Sorting does not rely on the inter-lingual document similarities at all, we hope that its supervised version is robust to noisy alignments.

Let X and Y be the TFIDF weighted term $\times$ document matrices in both the languages and let K_x and K_y be their linear dot product kernel matrices, *i.e.* , $K_x = X^T X$ and $K_y = Y^T Y$. Let $\Pi \in \{0,1\}^{m \times n}$ denote the permutation matrix which captures the alignment between documents of different languages, *i.e.* $\pi_{ij} = 1$ indicates documents x_i and y_j are aligned. Then Kernelized Sorting formulates Π as the solution of the following optimization problem (Gretton et al., 2005):

$$\arg\max_{\Pi} \; \mathrm{tr}(K_x \Pi K_y \Pi^T) \quad (2)$$
$$= \arg\max_{\Pi} \; \mathrm{tr}(X^T X \; \Pi \; Y^T Y \; \Pi^T) \quad (3)$$

In our supervised version of Kernelized Sorting, we fix the permutation matrix (to say $\hat{\Pi}$) and modify the kernel matrices K_x and K_y so that the objective function is maximized for the given permutation. Specifically, we find a mapping for each language, such that when the documents are projected into their common subspaces they are more likely to respect the alignment given by $\hat{\Pi}$. Subsequently, the test documents are also projected into the common subspace and we return the nearest neighbors as the aligned pairs.

Let U and V be the mappings for the required subspace in both the languages, then we want to solve the following optimization problem:

$$\arg\max_{U,V} \; \mathrm{tr}(X^T U U^T X \; \hat{\Pi} \; Y^T V V^T Y \; \hat{\Pi}^T)$$
$$\text{s.t.} \;\; U^T U = I \;\&\; V^T V = I \quad (4)$$

where I is an identity matrix of appropriate size. For brevity, let C_{xy} denote the cross-covariance matrix (*i.e.* $C_{xy} = X\hat{\Pi}Y^T$) then the above objective function becomes:

$$\arg\max_{U,V} \; \mathrm{tr}(U U^T C_{xy} V V^T C_{xy}^T)$$
$$\text{s.t.} \;\; U^T U = I \;\&\; V^T V = I \quad (5)$$

We have used the cyclic property of the trace function while rewriting Eq. 4 to Eq. 5. We use alternative maximization to solve for the unknowns. Fixing V (to say V_0), rewriting the objective function using the cyclic property of the trace function, forming the Lagrangian and setting its derivative to zero results in the following solution:

$$C_{xy} V_0 V_0^T C_{xy}^T \, U = \lambda_u \, U \quad (6)$$

For the initial iteration, we can substitute $V_0 V_0^T$ as identity matrix which leaves the kernel matrix unchanged. Similarly, fixing U (to U_0) and solving the optimization problem for V results:

$$C_{xy}^T U_0 U_0^T C_{xy} \, V = \lambda_v \, V \quad (7)$$

In the special case where both $V_0 V_0^T$ and $U_0 U_0^T$ are identity matrices, the above equations reduce to $C_{xy} C_{xy}^T \; U = \lambda_u \; U$ and $C_{xy}^T C_{xy} \; V = \lambda_v \; V$. In this particular case, we can simultaneously solve for both U and V using Singular Value Decomposition (SVD) as:

$$USV^T = C_{xy} \tag{8}$$

So for the first iteration, we do the SVD of the cross-covariance matrix and get the mappings. For the subsequent iterations, we use the mappings found by the previous iteration, as U_0 and V_0, and solve Eqs. 6 and 7 alternatively.

2.3 Summary

In this section, we describe our procedure to recover document alignments. We first convert documents into bag of dictionary entries representation (Sec. 2.1). Then we solve the optimization problem in Eq. 1 to get the initial candidate alignments. We use the LEMON[3] graph library to solve the min-cost flow problem. This step gives us the π_{ij} values for every cross-lingual document pair. We use them to form a relaxed permutation matrix ($\hat{\Pi}$) which is, subsequently, used to find the mappings (U and V) for the documents of both the languages (*i.e.* solving Eq. 8). We use these mappings to project both source and target language documents into the common subspace and then solve the bipartite matching problem to recover the alignment.

3 Experiments

For evaluation, we choose 2500 aligned document pairs from Wikipedia in English-Spanish and English-German language pairs. For both the data sets, we consider only words that occurred more than once in at least five documents. Of the words that meet the frequency criterion, we choose the most frequent 2000 words for English-Spanish data set. But, because of the compound word phenomenon of German, we retain all the frequent words for English-German data set. Subsequently we convert the documents into TFIDF weighted vectors. The bilingual dictionaries for both the language pairs are generated by running Giza++ (Och and Ney, 2003) on the Europarl data (Koehn, 2005).

[3]https://lemon.cs.elte.hu/trac/lemon

	En – Es	En – De
Word-by-Word	0.597	0.564
CCA (λ = 0.3)	0.627	0.485
CCA (λ = 0.5)	0.628	0.486
CCA (λ = 0.8)	0.637	0.487
OPCA	**0.688**	0.530
Ours (C = 0.6)	0.67	**0.604**
Ours (C = 1.0)	0.658	0.590

Table 1: Accuracy of different approaches on the Wikipedia documents in English-Spanish and English-German language pairs. For CCA, we regularize the within language covariance matrices as $(1-\lambda)XX^T+\lambda I$ and the regularization parameter λ value is also shown.

We follow the process described in Sec. 2.3 to recover the document alignment for our method.

We compare our approach with a dictionary based approach, such as word-by-word translation, and supervised approaches, such as CCA (Vinokourov et al., 2003; Hotelling, 1936) and OPCA (Platt et al., 2010). Word-by-word translation and our approach use bilingual dictionary while CCA and OPCA use a training corpus of aligned documents. Since the bilingual dictionary is learnt from Europarl data set, for a fair comparison, we train supervised approaches on 3000 document pairs from Europarl data set. To prevent CCA from overfitting to the training domain, we regularize it heavily. For OPCA, we use a regularization parameter of 0.1 as suggested by Platt et al. (2010). For all the systems, we construct a bipartite graph between the documents of different languages, with edge weight being the cross-lingual similarity given by the respective method and then find maximal matching (Jonker and Volgenant, 1987). We report the accuracy of the recovered alignment.

Table 1 shows accuracies of different methods on both Spanish and German data sets. For comparison purposes, we trained and tested CCA on documents from same domain (Wikipedia). It achieves 75% and 62% accuracies for the two data sets respectively but, as expected, it performed poorly when trained on Europarl articles. On the English-German data set, a simple word-by-word translation performed better than CCA and OPCA. For both the language pairs, our model performed better than word-by-word translation method and competitively with the

supervised approaches. Note that our method does not use any training data.

We also experimented with few values of the parameter C for the min-cost flow problem (Eq. 1). As noted previously, setting $C = 1$ will reduce the problem into a linear assignment problem. From the results, we see that solving a relaxed version of the problem gives better accuracies but the improvements are marginal (especially for English-German).

4 Discussion

For both language pairs, the accuracy of the first stage of our approach (Sec. 2.1) is almost same as that of word-by-word translation system. Thus, the improved performance of our system compared to word-by-word translation shows the effectiveness of the supervised Kernelized sorting.

The solution of our supervised Kernelized sorting (Eq. 8) resembles Latent Semantic Indexing (Deerwester, 1988). Except, we use a cross-covariance matrix instead of a term $\times$ document matrix. Efficient algorithms exist for solving SVD on arbitrarily large matrices, which makes our approach scalable to large data sets (Warmuth and Kuzmin, 2006). After solving Eq. 8, the mappings U and V can be improved by iteratively solving the Eqs. 6 and 7 respectively. But it leads the mappings to fit the noisy alignments exactly, so in this paper we stop after solving the SVD problem.

The extension of our approach to the situation with different number of documents on each side is straight forward. The only thing that changes is the way we compute alignment after finding the projection directions. In this case, the input to the bipartite matching problem is modified by adding dummy documents to the language that has fewer documents and assigning a very high score to edges that connect to the dummy documents.

5 Conclusion

In this paper we have presented an approach to recover document alignments from a comparable corpora using a bilingual dictionary. First, we use the bilingual dictionary to find a set of candidate noisy alignments. These noisy alignments are then fed into supervised Kernelized Sorting, which learns to modify within language document similarities to respect the given alignments.

Our approach exploits two complimentary information sources to recover a better alignment. The first step uses cross-lingual cues available in the form of a bilingual dictionary and the latter step exploits document structure captured in terms of within language document similarities. Experimental results show that our approach performs better than dictionary based approaches such as a word-by-word translation and is also competitive with supervised approaches like CCA and OPCA.

References

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.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Uncertainty in Artificial Intelligence*.

Scott Deerwester. 1988. Improving Information Retrieval with Latent Semantic Indexing. In Christine L. Borgman and Edward Y. H. Pai, editors, *Proceedings of the 51st ASIS Annual Meeting (ASIS '88)*, volume 25, Atlanta, Georgia, October. American Society for Information Science.

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, and Ming Zhou. 2008. Using english information in non-english web search. In *iNEWS '08: Proceeding of the 2nd ACM workshop on Improving non english web searching*, pages 17–24, New York, NY, USA. ACM.

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.

Arthur Gretton, Arthur Gretton, Olivier Bousquet, Olivier Bousquet, Er Smola, Bernhard Schlkopf, and Bernhard Schlkopf. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of Algorithmic Learning Theory*, pages 63–77. Springer-Verlag.

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.

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.

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

Jagadeesh Jagaralmudi, Seth Juarez, and Hal Daumé III. 2010. Kernelized sorting for natural language processing. In *Proceedings of AAAI Conference on Artificial Intelligence*.

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.

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 *MT Summit*.

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.

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.

Ari Pirkola, Turid Hedlund, Heikki Keskustalo, and Kalervo Jrvelin. 2001. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval*, 4:209–230.

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.

Novi Quadrianto, Le Song, and Alex J. Smola. 2009. Kernelized sorting. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1289–1296.

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.

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.

K. Ahuja Ravindra, L. Magnanti Thomas, and B. Orlin James. 1993. Network flows: Theory, algorithms, and applications.

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.

Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2009. Mint: A method for effective and scalable mining of named entity transliterations from large comparable corpora. In *EACL*, pages 799–807. The Association for Computer Linguistics.

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.

Manfred K. Warmuth and Dima Kuzmin. 2006. Randomized pca algorithms with regret bounds that are logarithmic in the dimension. In *Neural Information Processing Systems*, pages 1481–1488.