Towards Learning Coupled Representations for Cross-Lingual Information Retrieval

Rishabh Mehrotra[*]
Computer Science & Information Systems
BITS Pilani, India
erishabh@gmail.com

Dat Chu
Computational Biomedicine Lab
Dept. of Computer Science
University of Houston

Syed Aqueel Haider
SAP Labs, India
syed.aqueel.haider@sap.com

Ioannis A. Kakadiaris
Computational Biomedicine Lab
Dept. of Computer Science
University of Houston

Abstract

We explore the use of dictionary-based approaches for cross-lingual information retrieval tasks and propose a novel Coupled Dictionary Learning (CDL) algorithm to learn two separate representations simultaneously for documents in a parallel corpus alongside learning mappings from one representation to the other. We evaluate the performance of the proposed algorithm for the task of comparable document retrieval and compare with existing baselines.

1 Introduction

Automatic text understanding has been an unsolved research problem for many years. The challenges arise from the dynamic and diverging nature of human languages, which ultimately leads to many varieties of natural language. These variations range from the individual level, to regional and social dialects, and up to seemingly separate languages and language families. However, in recent years there have been considerable achievements in data-driven approaches to computational linguistics, which exploit the redundancy in the encoded information and the structures used. Most of these approaches are not specific to a particular language and are capable of finding the commonalities across languages. Representing documents by vectors that are independent of languages enhances the performance of cross-lingual tasks such as *comparable document retrieval* and *mate retrieval*. We address the problem of learning document representations which enable one to perform improved cross-lingual document retrieval.

There exist essentially two main paradigms to tackle the task of retrieving documents across languages. Firstly, the translation-based approaches, which rely either on a translation of documents or queries with the translation of queries usually done based on bilingual dictionaries which may not be always available for all sets of language pairs. The second paradigm involves mapping of queries and documents into a multilingual space [1][2] in which similarity between queries and documents can be computed uniformly across languages. A single multilingual space might not be best representative of the document representations while learning separate representations but in a coupled fashion might improve retrieval performance.

In this paper, we explore the use of dictionary-based approaches to solve the task of cross-lingual information retrieval by proposing a new dictionary learning algorithm (CDL: Coupled

[*] This work was conducted while the first author was an intern at the Computational Biomedicine Lab at University of Houston.

Dictionary Learning) for learning a pair of coupled dictionaries representing basis atoms in a pair of languages, alongside learning two separate mapping functions which help in transforming representations learnt in one language to the other. Such transformations are necessary for the task of finding similar documents in a different language and hence find application in various cross-lingual information retrieval tasks. We present an optimization procedure that iterates between two objectives and uses the K-SVD [3] formulation to efficiently compute the parameters involved. We evaluate our algorithm on the task of cross-lingual comparable document retrieval and compare our results with existing approaches. Our work is quite different from many pioneering studies on Cross-Lingual Information Retrieval as our proposed algorithm simultaneously learns coupled representations and the corresponding mapping functions alongside making use of unlabelled data to improve retrieval performance. The proposed CDL algorithm (2.2) and details of the iterative optimization (2.3) technique are discussed in section 2 followed by the its application to Cross-lingual information retrieval in section 3. Section 4 presents the results and concludes.

2 Coupled Dictionary Learning

The linear decomposition of a signal using few atoms of a *learned* dictionary has led to state-of-the-art performance in many computer vision and pattern recognition tasks. Recently it has been shown that learning dictionary based representations to model text corpora help in improving classification performance [4] as well as learning hierarchies of topics [5]. In this paper, we present a Coupled Dictionary Learning (CDL) algorithm 2.2 which simultaneously learns a pair of dictionaries and a pair of mapping functions to solve cross-lingual information retrieval problems. Specifically targeting the domain of resource-scarce languages, we propose the use of Coupled Dictionary Learning algorithm for cross-lingual document representation wherein the dictionary pair can well characterize the corpora of the two languages while the mapping function can reveal the intrinsic relationship between the language pair.

2.1 Problem Formulation

The cross-lingual document representation problem can be formulated as: given parallel corpora of a language pair $\langle L_1, L_2 \rangle$, can we learn document representations in each of these languages (Y_1 and Y_2) and their corresponding mappings ($T_{Y_1 \to Y_2}$ and $T_{Y_2 \to Y_1}$) so as to perform well in the challenging task of retrieving documents to queries in other languages. Since we are working with parallel corpora in our setting, it is reasonable to assume that there exists a latent space where these representations could be mapped to each other. 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 [6][2][7] or discriminative approaches based on variants of Principal Component Analysis and Canonical Correlation Analysis (CCA) [8]. Unlike existing methods, our framework makes use of abundantly available *unlabelled* data in each of the languages and learns meaningful intermediate representations and concepts which better capture the variations in naturally occurring data. Using this representation learnt in unsupervised fashion as initialization of the respective dictionaries, our algorithm then couples the learning of the two dictionaries and alongside learns mappings from each of the representation to the other. The unavailability of data in a resource-scarce language motivates the use of this mapping to preform computations in the transformed representation.

2.2 CDL Algorithm

We denote by $Y_1 \in R^{n \times N}$ and $Y_2 \in R^{n \times N}$ the training datasets formed by documents in the parallel corpora of the two languages. The corresponding dictionaries are notated as $D_1 \in R^{n \times K}$ and $D_1 \in R^{n \times K}$ with the mapping functions $T_{Y_1 \to Y_2} \in R^{K \times K}$ and $T_{Y_2 \to Y_1} \in R^{K \times K}$ with K being the number of dictionary atoms. Our framework is based on the Semi-Coupled Dictionary algorithm proposed in Wang et al[8]. We propose to minimize the following dictionary learning objective:

$$\begin{aligned}\langle D_1, D_2, X_1, X_2, T_{Y_1 \to Y_2} T_{Y_2 \to Y_1} \rangle \;&= \\ argmin_{\{D_1, D_2, T_{Y_1 \to Y_2}, T_{Y_2 \to Y_1}\}} \parallel Y_1 - D_1 X_1 \parallel_2^2 &+ \parallel Y_2 - D_2 X_2 \parallel_2^2 + \; \alpha \parallel X_2 - T_{Y_1 \to Y_2} X_1 \parallel_2^2 \\ &+ \; \beta \parallel X_1 - T_{Y_2 \to Y_1} X_2 \parallel_2^2\end{aligned}$$

s.t. $\|x_1^i\|_0 \leq T$ and $\|x_2^i\|_0 \leq T \;\; \forall i$, where $X_1 = \left[x_1^1, x_2^1, ..., x_n^1\right] \in R^{K \times N}$ are the sparse codes of the input data of language L_1 and $X_2 = \left[x_1^2, x_2^2, ..., x_n^2\right] \in R^{K \times N}$ are the sparse codes of the input data of language L_2 and T is the sparsity constraint factor.

The term $\| \ Y_i - D_i X_i \ \|_2^2$ for $i \in \{1, 2\}$ represents the reconstruction error for documents of both the languages which intuitively implies how well do the learnt representations represent the original documents. The parameters α and β control the relative contribution between reconstructive and mapping regularizations. By the term $\| \ X_2 - T_{Y_1 \to Y_2} X_1 \ \|_2^2$, we intend to minimize the mapping error between the transformed sparse codes of document in language L_1 and the corresponding document in L_2 while by $\| \ X_1 - T_{Y_2 \to Y_1} X_2 \ \|_2^2$ we intend to minimize the mapping error between the transformed sparse codes of documents in language L_1 and the corresponding document in L_1. This is the main contribution of our paper as we believe that if both the resource-scarce then ideally we should penalize errors in both the mapping functions $T_{Y_1 \to Y_2}$ and $T_{Y_2 \to Y_1}$.

When both the languages are resource-scarce, one might choose a language in which unlabelled data are more readily available to compute the initialized dictionary. Hence we might want to transform the document from its original language to the other language to perform retrieval tasks. This is our main motivation to have two separate transformation functions instead of a single one as in doing so we penalize the mapping errors in both the transforms and hence get optimized transforms for both the languages. Note that in the proposed model, the coding coefficients of X_1 and X_2 are related by the mapping functions $T_{Y_1 \to Y_2}$ and $T_{Y_2 \to Y_1}$ using which we could transform a document representation in language L_1 to its corresponding representation in L_2 and vice-versa.

2.3 Optimization

We use efficient K-SVD algorithm to find the optimal solution for all parameter simultaneously. This is quite different from the original approach as adopted in [8]. Since the objective is not jointly convex in all the parameters, Wang et al[8] use iterative algorithm to alternately optimize the parameters. Instead of following that approach, we iterate between the representations of the two languages using K-SVD to find the optimal solutions for all parameters. With the initialization of the dictionary pairs D_1 and D_2, the mapping functions $T_{Y_1 \to Y_2}$ and $T_{Y_2 \to Y_1}$, we iterate between the solutions of the following two equations:

$$\langle D_1^N, X_1 \rangle = argmin_{\{D_1^N, X_1\}} \| \ Y_1^N - D_1^N X_1 \ \|_2^2$$
$$\langle D_2^N, X_2 \rangle = argmin_{\{D_2^N, X_2\}} \| \ Y_2^N - D_2^N X_2 \ \|_2^2$$

s.t. $\forall i, \ \|x_1^i\|_0 \leq T$ and $\|x_2^i\|_0 \leq T$ where:

$$Y_1^N = \begin{pmatrix} Y_1 \\ \sqrt{\alpha} X_2 \end{pmatrix}; D_1^N = \begin{pmatrix} D_1 \\ \sqrt{\alpha} T_{Y_1 \to Y_2} \end{pmatrix}$$
$$Y_2^N = \begin{pmatrix} Y_2 \\ \sqrt{\beta} X_1 \end{pmatrix}; D_2^N = \begin{pmatrix} D_2 \\ \sqrt{\beta} T_{Y_2 \to Y_1} \end{pmatrix}$$

The matrices D_1^N and D_2^N are l_2-normalized column wise. The equations presented above are exactly the problem which K-SVD[3] solves. By solving the equations mentioned above, our algorithm learns a pair of dictionaries alongside learning mapping functions using which we can represent documents in both the languages and can map representations from one language to another so as to solve cross-lingual information retrieval tasks.

2.4 Dictionary Initialization

We need to initialize the parameters D_1, D_2, $T_{Y_1 \to Y_2}$ and $T_{Y_2 \to Y_1}$. To initialize D_i, $i \in \{1, 2\}$, several iterations of K-SVD are employed for each dictionary using unlabelled data from the corresponding language. This is in spirit of the Self-Taught Learning framework [9] where unlabelled data are used to learn an initial representation in an unsupervised manner. To the best of our knowledge, none of the existing approaches employed for Cross-Lingual information retrieval tasks make use of unlabelled data to improve the performance. Given the initialized dictionaries, K-SVD is used to compute the sparse codes X_i of the training data Y_i, $i \in \{1, 2\}$ which are then used to initialize the mapping parameters.

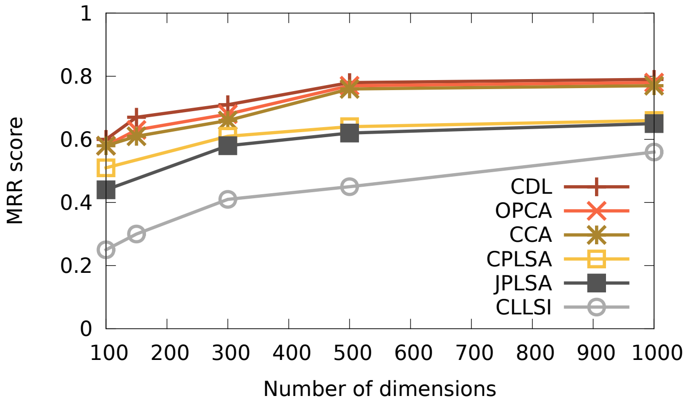

Figure 1: Comparison of MRR scores.

3 Cross-Lingual Document Retrieval

We obtain $D_i = [d_1^i, d_2^i, ..., d_k^i], i \in \{1, 2\}, T_{Y_1 \rightarrow Y_2} = [t_1^1, t_2^1, ..., t_k^1]$ and $T_{Y_2 \rightarrow Y_1} = [t_1^2, t_2^2, ..., t_k^2]$ by employing K-SVD algorithm in an iterative manner to the equations presented above. We cannot simply use these for testing since these are l_2-normalized in D_i^N jointly in our algorithm. Hence, we computed the desired dictionaries and mapping transformations as follows [10] :

$$D_1 = \left\{ \frac{d_1^1}{\| d_1 \|_2}, \frac{d_2^1}{\| d_2 \|_2}, ..., \frac{d_k^1}{\| d_k \|_2} \right\}; T_{Y_1 \rightarrow Y_2} = \left\{ \frac{t_1^1}{\| d_1 \|_2}, \frac{t_2^1}{\| d_2 \|_2}, ..., \frac{t_k^1}{\| d_k \|_2} \right\}$$
$$D_2 = \left\{ \frac{d_1^2}{\| d_1 \|_2}, \frac{d_2^2}{\| d_2 \|_2}, ..., \frac{d_k^2}{\| d_k \|_2} \right\}; T_{Y_2 \rightarrow Y_1} = \left\{ \frac{t_1^2}{\| d_1 \|_2}, \frac{t_2^2}{\| d_2 \|_2}, ..., \frac{t_k^2}{\| d_k \|_2} \right\}$$

For a test document z_j in language L_j $(j \in \{1, 2\})$ we compute its sparse representation x_j by solving the optimization problem: $x_j = argmin_{x_j}\{\| z_j - D_j x_j \|_2^2\}$ $s.t.$ $\forall i$ $\|x_j^i\|_0 \leq T$. For example, for the task of cross-lingual document retrieval, given a query document z_1 in language L_1 we find its representation x_1 and then use the mapping $T_{Y_1 \rightarrow Y_2}$ to transform this representation to obtain the corresponding representation in the target language domain where we compare it with all the documents using cosine based similarity score to find the most similar document from the corpus in the other language.

4 Results & Conclusion

In the cross-lingual document retrieval task, given a query document in one language, the goal is to find the most similar document from the corpus in another language. We followed the comparable document retrieval setting described in [1] and evaluated our algorithm on a subset of the Wikipedia dataset used in that paper. This data set consists of Wikipedia documents in two languages, English and Spanish. An article in English is paired with a Spanish article if they are identified as comparable across languages by the Wikipedia community. We evaluate the performance by using each English document as query against all documents in Spanish and vice versa; the results from the two directions are averaged. Performance is evaluated by the Mean Reciprocal Rank (MRR) of the true comparable. Our approach is compared with most methods studied in [1] (using the values presented there) including the best performing one: CL-LSI, OPCA, and CCA, JPLSA and CPLSA. Figure 1 summarizes our results.

In this work, we presented a novel algorithm for learning coupled representations for parallel documents belonging to a pair of languages alongside learning mapping functions from one representation to another. As a future extension of this work, we intend to develop dictionary learning algorithms aimed at learning discriminative translingual representations for classification tasks.

References

[1] P. Platt, K. Toutanova, and W. Yih, "Translingual document representations from discriminative projections," in *Proc. Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Association for Computational Linguistics,*, Stroudsburg, PA, USA, October 2010, pp. 251–261.

[2] J. Jagarlamudi, H. Daum III, and R. Udupa, "From bilingual dictionaries to interlingual document representations," in *Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*, Oregon, USA, June 2011, pp. 147–152.

[3] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, pp. 4311–4321, 2006.

[4] R. Mehrotra, R. Agrawal, and S. Haider, "Dictionary based sparse representation for domain adaptation," in *Proc. 21st ACM International Conference on Information and Knowledge Management*, Maui Hawaii, US, October 2012.

[5] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for sparse hierarchical dictionary learning," in *Proc. International Conference on Machine Learning*, Haifa, Israel, June 2010.

[6] J. Jagarlamudi and H. Daum, "Extracting multilingual topics from unaligned comparable corpora," *Advances in Information Retrieval*, pp. 444–456, 2010.

[7] D. Zhang, Q. Mei, and C. Zhai, "Cross-lingual latent topic extraction," in *Proc. 48th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, July 2010, pp. 1128–1137.

[8] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, Rhode Island, June 16-21 2012, pp. 2216–2223.

[9] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proc. 24th International Conference on Machine Learning*, Corvallis, Oregon, USA, June 20-24 2007, pp. 759–766.

[10] Z. Jiang, Z. Lin, and L. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *Proc. IEEE Computer Vision and Pattern Recognition*, Colorado Springs, USA, June 20-25 2011, pp. 1697–1704.