Learning Semantics with Deep Belief Network for Cross-Language Information Retrieval

Jungi Kim[1] *Jinseok Nam*[1] *Iryna Gurevych*[1,2]
(1) Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
(2) Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information
http://www.ukp.tu-darmstadt.de/
{kim,nam,gurevych}@ukp.informatik.tu-darmstadt.de

ABSTRACT
This paper introduces a cross-language information retrieval (CLIR) framework that combines the state-of-the-art keyword-based approach with a latent semantic-based retrieval model. To capture and analyze the hidden semantics in cross-lingual settings, we construct latent semantic models that map text in different languages into a shared semantic space. Our proposed framework consists of deep belief networks (DBN) for each language and we employ canonical correlation analysis (CCA) to construct a shared semantic space. We evaluated the proposed CLIR approach on a standard ad hoc CLIR dataset, and we show that the cross-lingual semantic analysis with DBN and CCA improves the state-of-the-art keyword-based CLIR performance.

KEYWORDS: Cross-language Information Retrieval, Ad Hoc Retrieval, Deep Learning, Deep Belief Network, Canonical Correlation Analysis, Wikipedia, CLEF.

1 Introduction

The state-of-the-art information retrieval (IR) systems of today rely on keyword matching, which suffers from the term mismatch problem. To this end, various techniques such as pseudo-relevance feedback and knowledge-based query expansion have been developed. More recently, a number of semantic analysis approaches such as word sense disambiguation (WSD), latent semantic indexing (LSI), latent dirichlet allocation (LDA), and explicit semantic analysis (ESA) have been utilized in IR (Wolf et al., 2010; Dumais et al., 1996; Vulić et al., 2011; Egozi et al., 2011).

The retrieval task becomes more difficult in the settings of cross-language information retrieval (CLIR), because of additional uncertainty introduced in the cross-lingual matching process. This paper introduces a CLIR framework that combines the state-of-the-art keyword-based approach with a latent semantic-based retrieval model (Fig. 1). To capture and analyze the hidden semantics of source language queries and documents in the target language, we construct latent semantic analysis models that map the texts in the source and the target languages into a shared semantic space, in which the similarities of a query and documents are measured. In addition to the traditional keyword-based CLIR system, our proposed framework consists of deep belief network (DBN)-based semantic analysis models for each language and a canonical correlation analysis (CCA) model for inter-lingual similarity computation. The DBN and the CCA models are trained on a large-scale comparable corpus and use low dimension vectors to represent the semantics of texts. The proposed approach is evaluated on a standard ad hoc CLIR dataset from CLEF workshop, with English as the source language and German as the target language.

2 Related Work

Deep learning is a machine learning approach that utilizes multiple layers of learners for modeling complex and abstract representations of input data. A recent introduction of an efficient deep learning architecture (Hinton et al., 2006) has contributed to its applicability to real world problems. There have been several uses of deep architectures in IR tasks such as mate retrieval (Salakhutdinov and Hinton, 2007; Ranzato and Szummer, 2008) and multimedia retrieval (Hörster and Lienhart, 2008; Krizhevsky and Hinton, 2011). These works find that high-level abstractions through deep networks achieve higher generalization than probabilistic topic models (Blei et al., 2003; Deerwester et al., 1990) in terms of unseen data.

Semantic analysis approaches for CLIR have used the notion of shared semantic space to represent and analyze the semantics across languages. In cross-lingual LSI (CL-LSI) (Dumais et al., 1996), the cross-lingual problem is translated to a monolingual one by merging comparable documents together. Polylingual topic model (Mimno et al., 2009) finds aligned semantics across languages, expanding the notion of the generative process of documents into a multilingual one. In Vulić et al. (2011), LDA-based interlingual topics are utilized in a language modeling approach to CLIR. These cross-lingual latent topic models are extended from monolingual models by blindly concatenating comparable document pairs. Therefore, they suffer from

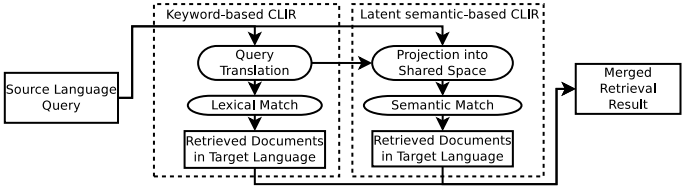


Figure 1: Proposed CLIR framework that combines both the keyword- and the latent semantic-based retrieval models.

reduced feature representations for each language because vocabulary space must be shared across languages.

Another line of research in discovering implicit semantic representation is utilizing CCA (Hotelling, 1936). CCA learns correlations between a pair of comparable representations and projects them into a shared space such that the correlation between the two spaces is maximized. Its recent applications include cross-lingual text analysis (Vinokourov et al., 2002; Hardoon et al., 2004; Li and Shawe-Taylor, 2007) and multi-modal learning (Rasiwasia et al., 2010; Ngiam et al., 2011). The advantage of CCA is that it can easily be generalized to incorporate multiple languages or modalities. Also, it has been shown to out-perform previously known state-of-the-art approaches such as probabilistic LSA and CL-CSI on the cross-lingual mate retrieval task (Platt et al., 2010). However, applying CCA on lexical representation of texts fails to capture complex and abstract relations, because CCA optimizes only on linear correlations.

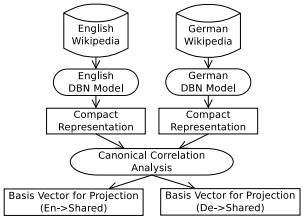
In contrast to these latent semantic methods, CL-ESA (Potthast et al., 2008; Cimiano et al., 2009) exploits explicit concepts to represent the semantics of texts. ESA methods assume that Wikipedia articles contain distinct topics and a set of Wiki articles can be used as concept features (Gabrilovich and Markovitch, 2007). CL-ESA additionally utilizes interlingual mappings in Wikipedia to find comparable articles, which are considered to be the same concept.

3 Overall approach

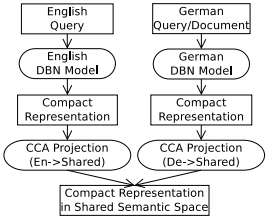
We propose a CLIR framework that combines the state-of-the-art keyword matching-based CLIR approach with a DBN-based retrieval model (Fig. 1). The intuition is to exploit the semantics of texts by measuring the similarities of a query and documents in addition to the keyword matching-based similarities. In the following sections, we explain the latent semantic-based (Sec. 3.1) and the keyword-based (Sec. 3.2) CLIR approaches and how the two approaches are merged (Sec. 3.3).

3.1 Deep Belief Network-based CLIR

The task of semantic-based CLIR is to discover a subset of documents in the target language that coincides with a query in the source language in a semantic space. This work utilizes a three-step approach as a latent semantic-based CLIR. First, DBN maps a lexical representation of a document into a latent semantic space (*semantic analysis* step). In the *semantic transformation* step, CCA maps the semantic representations of queries and documents into a shared semantic space. Finally, the *semantic matching* step retrieves relevant documents of a query using a distance metric in the shared semantic space.



(a) Training DBN and CCA models with a comparable corpus



(b) Mapping query and document into a shared semantic space

Figure 2: Overview of the DBN-based CLIR Framework: offline construction of models (a) and online inference processes (b).

To analyze the semantics of German queries and English documents, we train DBN models on German and English Wikipedia (Fig. 2a) and utilize the compact code resulting from the DBN models as semantic representations (Fig. 2b). Because the outputs from the German and English DBN models are from two different semantic spaces, we train a Canonical Correlation Analysis model (Fig. 2a) to map the German and English compact codes into a shared semantic space (Fig. 2b). To train a CCA model, we create a comparable corpus from Wikipedia using interlingual links between the English and German Wiki articles (Fig. 4).

The shared semantic space can be considered as an intermediate language representation, and the mapping process of the German and the English compact codes into the shared semantic space can be considered as the cross-lingual matching process of the two languages. Once mapped into a shared semantic space, the task of measuring the similarities between the semantic representations becomes trivial.

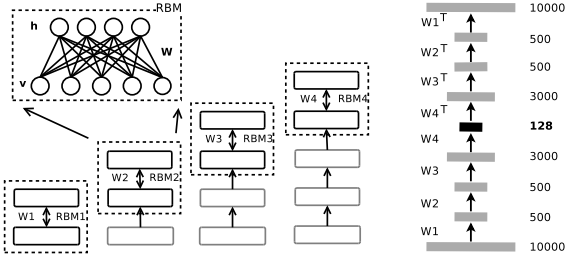
3.1.1 Semantic analysis with Deep Belief Network

To construct a DBN model, we follow the architecture of the model introduced in (Salakhutdinov and Hinton, 2007). DBN consists of stacked Restricted Boltzmann Machines (RBMs). Each RBM layer is trained in a greedy layer-wise manner (pretraining) and parameters of the entire model are adjusted (fine-tuning).

Pretraining An RBM(Smolensky, 1986) is a special form of Boltzmann Machine with bipartite connectivity constraints in which a set of visible units **v** are connected via symmetric weights **W** to a set of hidden units **h**. In the training step, the edge weights between visible and hidden units are updated iteratively given the input data. In the inference step, when the visible units are activated according to the input data, hidden units activated by the internal stochastic process is considered the hidden representation of the input data.

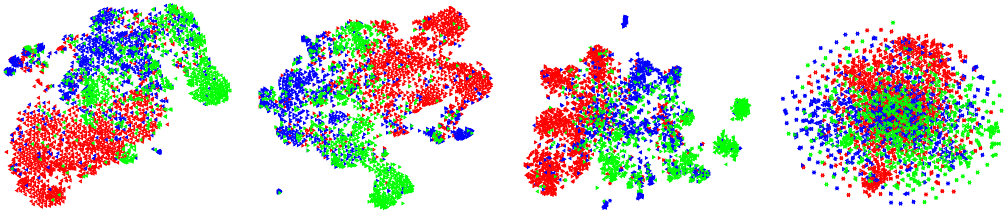
The bottom-most RBM accepts as input bag-of-word vector representations of texts processed with the replicated softmax model (RSM) (Salakhutdinov and Hinton, 2009). Upper RBMs are inputted with outputs from the binary RBM in the layer below.

An RBM is frequently explained using the *energy*-based analogy borrowed from Physics. The marginal distribution over an input vector **v** in a form of energy-based model (LeCun et al., 2006) is formulated as $p(\mathbf{v}) = \sum_{\mathbf{h}} \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z}$ where $Z = \sum_{\mathbf{u},\mathbf{g}} e^{-E(\mathbf{u},\mathbf{g})}$ is a normalization factor and $E(\mathbf{v},\mathbf{h}) = -\sum_{i,j} v_i h_j W_{ij} - \sum_i c_i v_i - \alpha \sum_j b_j h_j$ is an energy function of the RBM's state. c_i and b_j are biases for v_i and h_j, and α denotes a scale factor for hidden units and biases. We set α to the document length for discrete valued visible units, and α is set to 1 for binary valued visible units. The conditional distribution of **v** is $p(h_j = 1|\mathbf{v}) = \sigma\left(b_j + \sum_i v_i W_{ij}\right)$. In case of RSM, the



(a) Greedy layer-wise pretraining (b) Fine-tuning

Figure 3: Two-phase learning steps of DBN. Pretraining initializes RBM parameters from bottom to top (a), and fine-tuning optimizes them globally (b).



(a) English semantic space (b) German semantic space (c) Shared semantic space (DBN-CCA) (d) Shared lexical space (CCA)

Figure 4: Semantic spaces of English and German DBN models trained on Wikipedia and CCA model. Colors indicate the category of Wiki articles; Red indicates Art, Green is Science, and Blue is Education. The datasets are visualized with a variant of the Stochastic Neighbor Embedding technique (van der Maaten and Hinton, 2008).

conditional distribution of $\mathbf{h}$ is $p(v_i = 1|\mathbf{h}) = \frac{\exp(c_i + \sum_j h_j w_{ij})}{\sum_{k=1} \exp(c_k + \sum_j h_j w_{kj})}$, and in case of binary RBMs, it is $p(v_i = 1|\mathbf{h}) = \sigma(c_i + \sum_j h_j W_{ij})$.

Fine-tuning At the fine-tuning step, pretrained parameters of RBMs are fixed and stacked RBMs are unrolled as a deep autoencoder (Fig. 3b), which learns an identity function with constraints on a small number of hidden units. The constraints enable the autoencoder to find useful hidden representations. The deep autoencoder is trained by backpropagating reconstruction error at the top layer. The activation probability of the central layer (filled rectangle in Fig. 3b, with dimension of 128) is the most abstract and complex representation of the input text, and it is utilized as the latent semantic feature for the text in our work.

3.1.2 Semantic transformation with Canonical Correlation Analysis

CCA (Hotelling, 1936) is a method for discovering a subspace where the correlations of input spaces are maximized when linearly combined. Consider a pair of documents $\{(D_{x1}, D_{y1}), (D_{x2}, D_{y2}), \ldots, (D_{xN}, D_{yN})\} \in \mathbf{D}$ where $D_{xi} \in \mathbf{D}_x$ and $D_{yi} \in \mathbf{D}_y$ and a function $f(x)$ that maps a document into its semantic space. CCA seeks basis vectors w_x and w_y that maximize cross-lingual correlation $\rho = \max_{w_x, w_y} \frac{<w_x f(\mathbf{D}_x), w_y f(\mathbf{D}_y)>}{\|w_x f(\mathbf{D}_x)\| \|w_y f(\mathbf{D}_y)\|}$. We solve this optimization task as a generalized eigenvalue problem using a publicly available toolbox(Hardoon et al., 2004).[1] Figures 4a and 4b illustrate the DBN models trained on the English and German Wikipedia datasets, in which three categories (Art, Science, and Education) are shown for demonstration purposes.

3.1.3 Semantic matching with Cosine similarity

Given a semantic representation of an English query q_x and directions w_x and w_y for English semantics and German semantics, we use Cosine similarity $(f(q_x), f(d_{y_i}); w_x, w_y) = \frac{w_x f(q_x) \cdot w_y f(d_{y_i})}{\|w_x f(q_x)\| \|w_y (d_{y_i})\|}$ to produce the scores of German documents D_y. Note that CCA uses this measure to maximize the correlations of semantics over German and English Wiki articles, hence it is reasonable to use the same measure in the semantic matching phase.

[1] `http://www.davidroihardoon.com/Professional/Code.html`

Table 1: Statistics of the test collections. (T: Title, D: Description, N: Narrative, Cnt: Count)

Collection	Domain	Document			Topic			
		Lang	Cnt	Avg Len	Lang	Cnt	# Rel Doc	Avg Len (T/D/N)
AH 2001-2	News	DE	294,339	323.42	EN	100	4,021	3.38/8.32/18.1
AH 2003						60	1,825	3.83/8.18/17.7

3.2 Keyword matching-based CLIR

As a baseline CLIR framework, we employ a combination of a monolingual IR system with a commercial state-of-the-art MT system to translate English queries into German text.[2]

3.3 Merging Retrieval Results

Overall, there are three sets of retrieved documents: two from the DBN-based CLIR using the original query in the source language and its translation in the target language, and a set of retrieved documents from the keyword-based CLIR.

First, the two sets of retrieved documents from the DBN-based CLIR are linearly combined as:

$$Score_{DBN}(d_i,q) = \beta \cdot \frac{Score_{DBN}(d_i,q_{EN})}{\max_j\{Score_{DBN}(d_j,q_{EN})\}} + (1-\beta)\cdot \frac{Score_{DBN}(d_i,q'_{DE})}{\max_j\{Score_{DBN}(d_j,q'_{DE})\}} \quad (1)$$

in which, the two sets of document scores are first normalized with the maximum scores in each set of scores, and combined together with a ratio optimized on the development data.

Secondly, the document scores in the retrieval results from the BM25 and DBN are merged to produce the final retrieval results.

$$Score(d_i,q) = \lambda \cdot \frac{Score_{BM25}(d_i,q)}{\max_j\{Score_{BM25}(d_j,q)\}} + (1-\lambda)\cdot \frac{Score_{DBN}(d_i,q)}{\max_j\{Score_{DBN}(d_j,q)\}} \quad (2)$$

4 Experiment

Experimental setting We use standard newspaper ad-hoc retrieval datasets[3] for English to German CLIR experiments. General statistics of the test collections are presented in Table 1. As an evaluation measure, we use the mean average precision (MAP), which has been most widely used for evaluating IR systems.

Text preprocessing We obtained English and German comparable articles from Wikipedia dumps from October 07, 2011. For DBN-CLIR, term weights are evaluated with BM25 with an estimation to the nearest integer. After the conversion from texts to real number vectors, all pairs of documents satisfying $|D_{EN}| \geq 3 \times |D_{DE}|$ are dropped where $|D|$ represents the sum of feature values in document D. Out of 700,000 document pairs, 50,000 were randomly selected as training data.

Training DBNs We trained our DBNs with 4 hidden layers, 500-500-3000-128, meaning that 500 hidden units are at the first hidden layer, 500 units at the second, 3000 units at the third, and 128 at the highest level of hidden layer. Each RBM is trained with a mini-batch size of 100 training pairs for 50 epochs. The weights were updated by learning rate 0.1, weight decay 2×10^{-4} and momentum 0.9 except the first RBMs, in which the learning rate was set to 10^{-4}.

[2]Terrier (BM25 with Bo1, `http://terrier.org/`) and Google Translate (`http://translate.google.com/`)

[3]`http://www.clef-initiative.eu/track/series/`

Table 2: Mean average precision (MAP) of CLIR runs with different retrieval models. *%Chg* indicates the relative performance achieved compared to the baseline approach. Significant improvements (Student's paired t-test, $p<0.05$ and $p<0.01$) over the baseline approaches are marked with † and ‡.

Retrieval model	Topic Field					
	T		TD		TDN	
	MAP	*%Chg*	MAP	*%Chg*	MAP	*%Chg*
	AH 2001-2 (λ and β optimized on AH 2003)					
Monolingual (BM25)	24.13	–	28.82	–	30.92	–
Monolingual (BM25Bo1)	28.83	–	33.61	–	35.36	–
BM25	26.04	–	30.92	–	33.29	–
BM25+CCA	26.04	0.00	30.40	−1.68	33.29	0.00
BM25+DBN-CCA	26.32†	+1.07	31.24	+1.03	33.36	+0.21
BM25Bo1	28.97	–	34.64	–	37.35	–
BM25Bo1+CCA	28.82	−0.52	34.64	0.00	37.35	0.00
BM25Bo1+DBN-CCA	29.16	+0.65	35.07‡	+1.24	37.70	+0.94
	AH 2003 (λ and β optimized on AH 2001-2)					
Monolingual (BM25)	31.94	–	35.92	–	39.58	–
Monolingual (BM25Bo1)	39.25	–	40.13	–	44.73	–
BM25	30.40	–	36.01	–	36.43	–
BM25+CCA	30.40	0.00	35.93	−0.22	36.43	0.00
BM25+DBN-CCA	31.25‡	+2.80	36.62	+1.69	37.60†	+3.21
BM25Bo1	33.19	–	43.22	–	40.70	–
BM25Bo1+CCA	33.19	0.00	42.82	−0.93	40.70	0.00
BM25Bo1+DBN-CCA	33.26	+0.21	43.35	+0.30	40.89	+0.47

For fine-tuning phase, parameters of deep autoencoders were trained based on the L-BFGS optimization algorithm[4] with 5 line searches in each iteration.

CLIR results We performed a number of CLIR experiments with different retrieval models and parameter settings, such as the combinations of topic fields for querying (T, TD, and TDN) and whether or not query expansion (Bo1) is applied (Table 2). Performances of monolingual runs are also provided, which provide *soft* upper bounds to the cross-lingual runs. The baseline in our experiments is the state-of-the-art keyword-based CLIR (BM25). Also, we compare DBN-CCA with CCA, which has been shown to achieve top performances among state-of-the-art latent semantic approaches (Platt et al., 2010).

Our proposed approach utilizes a smoothing parameter λ and β to merge BM25 and DBN-based retrieved results (Eqs. 1 and 2). We use AH 2001-2 dataset for parameter estimation for AH 2003, and parameter estimation for AH 2001-2 is carried out on AH 2003.[5]

Table 2 reveals that merging DBN retrieved results with state-of-the-art CLIR approaches further improves the retrieval effectiveness; in all CLIR runs where DBN-CCA is utilized, we observe relative performance improvements in MAP in the range of $+0.21 \sim +3.21\%$.We also observe that some improvements are statistically significant. The improvements are consistent over the two test datasets as well as the combinations of topic fields, though utilization of query expansion decreases the effect of DBN, especially for the AH 2003 dataset.

CCA-based retrieval approaches resulted in negative results; in most runs, the optimal value for λ was 1.00, indicating that any portion of retrieval results from CCA only damages the overall performance.

[4]minFunc toolbox for Matlab `http://www.di.ens.fr/~mschmidt/Software/minFunc.html`

[5]For estimation of λ, we used an incremental step 0.05 from 0.0 to 1.0, and β, 0.00 to 1.00 with a step of 0.10.

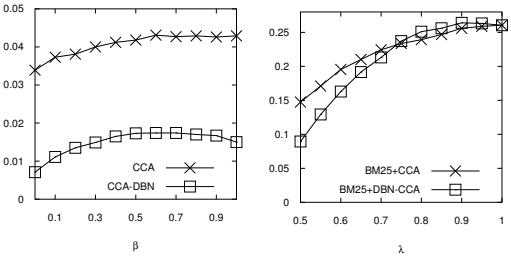


Figure 5: Comparing CLIR performances of CCA vs. DBN-CCA latent semantic methods on AH 2001-2 dataset with Title (T) topic field (x-axis: interpolation parameter, y-axis: MAP). In (b), β for the two runs are fixed to their optimal values (0.1 in both runs).

(a) Using latent semantic models alone ($\lambda = 0.0$)

(b) Using keyword and latent semantic fusion models (correspond to rows 4 and 5 in Table 2)

5 Discussion

Effect of utilizing DBN and CCA models in CLIR We conclude from the experimental results that a DBN-based semantic model helps better represent the query and documents and better matches across the language barrier. We observe, however, that its coverage is limited at its current implementation. This limitation is caused by a number of factors. First, our DBN models have lexicon sizes of only 10,000 terms in each language, which may lead to a lexical coverage problem. Secondly, our framework would not have any effect on topics where the lexical term mismatching problem is not severe. This explains the reduced effectiveness on experiments where query expansion is applied. On the brighter side, applying a DBN semantic model does not harm the overall retrieval performances, even when it is not effective.

CCA vs. DBN-CCA Fig. 4d shows visualization of CCA model trained on the bag-of-word representations of English and German Wikipedia articles with a dimension of 10,000. Compared to the output of DBN-CCA (Fig. 4c), the articles of same topics are scattered over a wider area sporadically and the clusters of topics are less apparent. For capturing underlying semantics, it is clear that DBN-CCA approach is superior over CCA-only method.

We observed in our post-experiment analysis that CCA and DBN-CCA have different roles in retrieval; CCA outperforms DBN-CCA when it is used alone in the retrieval process (Fig. 5). However, when combined with a keyword-based retrieval model, CCA only harms the overall performance while DBN-CCA marginally improves the performance. We attribute this to the fact that CCA and keyword-based IR both operate on the lexical level. The combination of DBN-CCA and keyword-based IR is however complementary because DBN-CCA captures topical similarities, introducing a additional information to the task.

Conclusion

This paper introduced a new latent semantic-based CLIR framework based on DBN and CCA. Though our proposed CLIR framework utilizes a relatively simple fusion approach, we showed that the cross-lingual semantic analysis with DBN and CCA improves the state-of-the-art keyword-based and CLIR performance.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806 and by the German Research Foundation under grant 798/1-5.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Cimiano, P., Schultz, A., Sizov, S., Sorg, P., and Staab, S. (2009). Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1513–1518.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Dumais, S., Landauer, T., and Littman, M. (1996). Automatic cross-language information retrieval using latent semantic indexing. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 16–23.

Egozi, O., Markovitch, S., and Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Trans. Inf. Syst.*, 29(2):8:1–8:34.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on artifical intelligence*, pages 1606–1611, Hyderabad, India.

Hardoon, D., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Hinton, G., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

Hörster, E. and Lienhart, R. (2008). Deep networks for image retrieval on large-scale databases. In *Proceeding of the 16th ACM international conference on multimedia*, page 643, New York, New York, USA. ACM Press.

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):pp. 321–377.

Krizhevsky, A. and Hinton, G. E. (2011). Using very deep autoencoders for content-based image retrieval. In *Proceedings of the 18th European Symposium On Artificial Neural Networks, Computational Intelligence and Machine Learning*. ESANN.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M. A., and Huang, F.-J. (2006). A Tutorial on Energy-Based Learning. In Bakir, G., Hofman, T., Schölkopf, B., Smola, A., and Taskar, B., editors, *Predicting Structured Data*. MIT Press.

Li, Y. and Shawe-Taylor, J. (2007). Advanced learning algorithms for cross-language patent retrieval and classification: Patent Processing. *Information Processing & Management*, 43(5).

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889, Singapore. Association for Computational Linguistics.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal Deep Learning. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 689—-696. ACM.

Platt, J., Toutanova, K., and Yih, W.-t. (2010). Translingual Document Representations from Discriminative Projections. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 251–261, Cambridge, MA. Association for Computational Linguistics.

Potthast, M., Stein, B., and Anderka, M. (2008). A wikipedia-based multilingual retrieval model. In Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., and White, R., editors, *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 522–530. Springer Berlin/Heidelberg.

Ranzato, M. A. and Szummer, M. (2008). Semi-supervised learning of compact document representations with deep networks. In *Proceedings of the 25th international conference on Machine learning*, pages 792–799, New York, New York, USA. ACM Press.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., and Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of the international conference on Multimedia*, pages 251–260, New York, New York, USA. ACM Press.

Salakhutdinov, R. and Hinton, G. (2009). Replicated softmax: an undirected topic model. *Advances in Neural Information Processing Systems*, 22:1607–1614.

Salakhutdinov, R. R. and Hinton, G. E. (2007). Semantic Hashing. In *Proceedings of the SIGIR Workshop on Information Retrieval and Applications of Graphical Models*, Amsterdam. Elsevier.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing*, chapter 6, pages 194–281. Dept. of Computer Science, University of Colorado, Boulder.

van der Maaten, L. and Hinton, G. E. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Vinokourov, A., Shawe-taylor, J., and Cristianini, N. (2002). Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis. In *Advances of Neural Information Processing Systems*, pages 1473–1480, Vancouver, British Columbia, Canada. MIT Press.

Vulić, I., Smet, W. D., and Moens, M.-F. (2011). Cross-Language Information Retrieval with Latent Topic Models Trained on a Comparable Corpus. In Mohamed-Salem, M.-V. O., Shaalan, K. F., Oroumchian, F., Shakery, A., and Khelalfa, H. M., editors, *Proceedings of the 7th Asia Information Retrieval Societies Conference*, volume 7097 of *Lecture Notes in Computer Science*, pages 37–48, Dubai, United Arab Emirates. Springer.

Wolf, E., Bernhard, D., and Gurevych, I. (2010). Combining probabilistic and translation-based models for information retrieval based on word sense annotations. In et al. (Eds.), C. P., editor, *CLEF 2009 Workshop, Part I*, volume 6241 of *Lecture Notes in Computer Science*, pages 120–127. Springer, Berlin/Heidelberg.