Continuous Space Models for CLIR

Parth Gupta[a,*], Rafael E. Banchs[b], Paolo Rosso[a]

[a]*PRHLT Research Center, Universitat Politècnica de València, Spain*
[b]*Institute for Infocomm Research, Singapore*

Abstract

We present and evaluate a novel technique for learning cross-lingual continuous space models to aid cross-language information retrieval (CLIR). Our model, which is referred to as external-data composition neural network (XCNN), is based on a composition function that is implemented on top of a deep neural network that provides a distributed learning framework. Different from most existing models, which rely only on available parallel data for training, our learning framework provides a natural way to exploit monolingual data and its associated relevance metadata for learning continuous space representations of language. Cross-language extensions of the obtained models can then be trained by using a small set of parallel data. This property is very helpful for resource-poor languages, therefore, we carry out experiments on the English-Hindi language pair. On the conducted comparative evaluation, the proposed model is shown to outperform state-of-the-art continuous space models with statistically significant margin on two different tasks: parallel sentence retrieval and ad-hoc retrieval.

Keywords:
cross-language information retrieval, latens space models

1. Introduction

Vector space models (VSM) and probabilistic information retrieval models provide a way to compare documents and queries by different means of

*Corresponding author
Email addresses: pgupta@dsic.upv.es (Parth Gupta), rembanchs@i2r.a-star.edu.sg (Rafael E. Banchs), prosso@dsic.upv.es (Paolo Rosso)

keyword matching. However, such lexical matching can be inaccurate due to the fact that the relevance is often expressed by different vocabularies in documents and queries. One of the major hurdles in comparing text in VSM is to deal with problems like *synonymy* and *polysemy*. Usually in vector space, the documents are composed of thousands of independent dimensions resulting in many meaningful associations between terms being neglected by the dimensional independence e.g. "small" and "petite" are not really independent/orthogonal. This problem is even more persistent in case of cross-language similarity estimation as the vocabulary overlap between language is little for languages sharing the same script and none for languages using different writing systems. There are models which try to handle this problem in the vector space *e.g.* pseudo relevance feedback (PRF) and explicit semantic analysis (ESA) (Xu and Croft, 1996; Gabrilovich and Markovitch, 2007). Other category of attempts to solve this problem comprises dimensionality reduction techniques often referred to as latent semantic models.

Latent semantic models map the high dimensional term vectors into a low dimensional abstract space referred to as latent space. There are broadly two categories of approaches: *i)* generative topic models, and *ii)* projection based models. Generative topic models, like latent dirichlet allocation (LDA), represent the high dimensional term vectors in a low-dimensional latent space of hidden topics. The projection based methods, like latent semantic analysis (LSA), learn a projection operator to map high-dimensional term vectors to low-dimensional latent space (Deerwester et al., 1990; Dumais et al., 1997; Platt et al., 2010; Yih et al., 2011). There exist cross-lingual variants of these models and discussed further in Sections 2 and 3. These models can be further categorised according to the objective function they optimise and the type of data they take in. Most of these models optimise an objective function which only loosely relates to the evaluation metric of the retrieval task and they leverage only parallel/comparable data to learn the joint latent space. This can prove to be a severe limitation for the resource-poor languages for which a large amount of bilingual data is not available.

In this paper, we propose a novel method to learn cross-lingual term associations in distributed manner to aid cross-language information retrieval (CLIR). In contrast to most of the existing models which rely only on the comparable/parallel data (now onward referred to as parallel data), our model takes in the external relevance signals such as the pseudo-relevant data to initialise the space monolingually and then, with the use of a small amount of parallel data, adjusts the parameters for different languages. There are a

few approaches which go beyond the use of only parallel data. The framework also allows the use of clickthrough data if available instead of pseudo-relevant data. Our model, differently from other models, optimises an objective function that is directly related to an evaluation metric for retrieval tasks such as cosine similarity. These two properties prove crucial for our model to outperform existing techniques in cross-lingual IR setting. We test our model on two different tasks of CLIR: parallel sentence retrieval and ad-hoc retrieval. The proposed model has the best performance in comparison to a number of strong baselines including machine translation based vector space models.

We present some related work in Section 2 and describe in detail the most relevant existing approaches in Section 3. The details of our approach are presented in Section 4. In Section 5, we present the experimental setup and results with analysis for the parallel sentence retrieval and ad-hoc retrieval tasks. Finally in Section 6, we draw conclusions.

2. Related Work

Latent semantic models such as the LSA are able to correspond queries and relevant documents at the semantic level where lexical matching often fails (Deerwester et al., 1990; Blei et al., 2003; Salakhutdinov and Hinton, 2009; Hinton and Salakhutdinov, 2009; Platt et al., 2010; Huang et al., 2013). These latent semantic models represent the text in a dense low-dimensional semantic space where the semantically similar text fragments would be closer to each other despite the fragments do not share any term. The semantic representation is learned through the patterns of terms occurring in similar contexts. LSA extracts a low rank Gaussian approximation of a document-term matrix by means of singular value decomposition (SVD) (Deerwester et al., 1990). More advanced approaches like probabilistic latent semantic analysis (PLSA) and latent dirichlet allocation (LDA) observe the distribution of latent topics for the given documents (Hofmann, 1999; Blei et al., 2003). Salakhutdinov and Hinton (2009) proposed an alternative approach to semantic modelling through the use of deep autoencoders. They showed such models would lead to more compact and superior representation of data compared to linear counterparts such as the LSA. Mikolov et al. (2013a) proposed a powerful technique to learn word embeddings known as word2vec. Gupta et al. (2016) have extended deep autoencoders to resolve lexical selection problem for machine translation. However, these models are trained to optimise an objective function which is only loosely related to the evalu-

ation metric of the retrieval task. To overcome this limitation, a new family of latent semantic models have emerged that exploits the clickthrough data for semantic modelling (Gao et al., 2010, 2011; Huang et al., 2013). These models take into account an explicit relevance signal in terms of the query and its clicked document.

Similar to cross-language text similarity, there are two broad approaches to CLIR: *i)* an off-the-self machine translation (MT) system is used to translate the data to the language of comparison followed by a standard IR technique e.g. TF-IDF or BM25, and *ii)* a cross-language latent semantic model can be applied to project the data into a low-dimensional translingual space where the texts can be compared. Though the MT based language normalisation can be highly accurate, the retrieval suffers from the already discussed issues of high-dimensional IR in Section 1. Moreover, MT can be very slow, limiting its use on large training datasets (Platt et al., 2010). In order to tackle the speed issue, the word-by-word translation models are used in which a translation dictionaries are learned from the parallel data with standard IR techniques (Ballesteros and Croft, 1996; Nie et al., 1999). There are also attempts to incorporate relevance feedback into estimating word-level translation probabilities which is also relevant to our work (Hiemstra et al., 2001). Alternatively, the cross-language latent semantic models provide a way to model cross-language term associations in the latent space. Such models include LSA based cross-language latent semantic analysis (CL-LSA) (Dumais et al., 1997) in which the document-term matrix is represented by concatenating the parallel data. Canonical correlation analysis (CCA) based methods find projections that maximises the correlation between the projected vectors of parallel data (Vinokourov et al., 2002). Generative models, such as the LDA, are used to represent bilingual data into hidden topical space (Mimno et al., 2009). Oriented principal component analysis (OPCA) introduces the noise covariance matrix and solves the generalised eigenvalue problem (Diamantaras and Kung, 1996; Platt et al., 2010). There are also a few dimensionality reduction techniques which are not based on matrix factorization. Deep bilingual autoencoders (BAE) are used to represent bilingual data in a low-dimensional joint space by optimising the reconstruction error (Lauly et al., 2014; Chandar A. P. et al., 2014; Gupta et al., 2014). Siamese neural network based S2Net learns discriminatively the projection matrix from the pairs of related and unrelated documents (Yih et al., 2011). Except for the S2Net method, which is closest to our work, all these models derive cross-language representations in an unsupervised manner by optimising an

objective function which only loosely related to the evaluation metric for the retrieval task. We review some of these models in detail in Section 3 and compare them to our proposed model in Section 5.

Another family of models for cross-language natural language processing applications, require advanced syntactic information in input such as the syntactic parse trees (Socher et al., 2012; Hermann and Blunsom, 2013). Similar models sometimes also require word-alignments during the training (Klementiev et al., 2012; Zou et al., 2013; Mikolov et al., 2013b). Such requirements limit the use of these approaches to resource fortunate languages.

3. State-of-the-art Baseline Systems

In this section, we review the state-of-the-art cross-language latent semantic models for information retrieval and machine translation based baseline systems.

3.1. Cross-Language Latent Semantic Indexing (CL-LSI)

CL-LSI performs singular value decomposition of document-term matrix D (Dumais et al., 1997). CL-LSI obtains the top k principal components of D that form the projection space in which documents can be compared on a semantic basis. The inherent idea is that semantically similar terms across languages (dimensions of D) will correspond to similar latent components. According to this, semantically similar documents will appear close to each other in the reduced comparison space.

This method is closely related to the eigenproblem, which is formulated as follows:

$$Cv_j = \lambda_j v_j, \tag{1}$$

where, λ_j is the j^{th} largest eigenvalue, v_j is corresponding eigenvector and C is correlation matrix ($D^T D$). In this setting, CL-LSI uses the top k eigenvectors for projection.

3.2. Oriented Principal Component Analysis (OPCA)

OPCA formulates the problem in a more structured way by introducing a noise component. It solves the generalised eigenproblem, which maximises the signal-to-noise ratio (Platt et al., 2010).

$$Sv_j = \lambda_j N v_j, \tag{2}$$

where, S is C-like matrix and N is covariance matrix of the differences among parallel documents which are considered noise.

Theoretically, OPCA tries to minimise the distance between the parallel documents at the same time of maximising the overall variance of the data. The parameters of OPCA are tuned according to Platt et al. (2010).

3.3. Bilingual Autoencoder (BAE)

Salakhutdinov and Hinton (2009) demonstrated that text representation learning by means of dimensionality reduction through deep autoencoders lead to superior performance compared to the conventional LSA approach. Deep autoencoders provide a non-linear generalisation of principal component analysis (PCA) through multi-layer architecture which help them to achieve better representation learning in more compact dimensionality (Hinton and Salakhutdinov, 2006). Deep autoencoders were extended to model cross-language data and are referred to as bilingual autoencoders (Gupta et al., 2014; Lauly et al., 2014; Chandar A. P. et al., 2014). These networks learn cross-language associations by optimising the reconstruction error of the cross-language data.

The building block of the autoencoder is the Restricted Boltzmann Machine (RBM). These deep networks are trained through a greedy layer-by-layer pretraining stage followed by a supervised fine-tuning. The structures of the network and the training architecture are shown in Fig. 1.

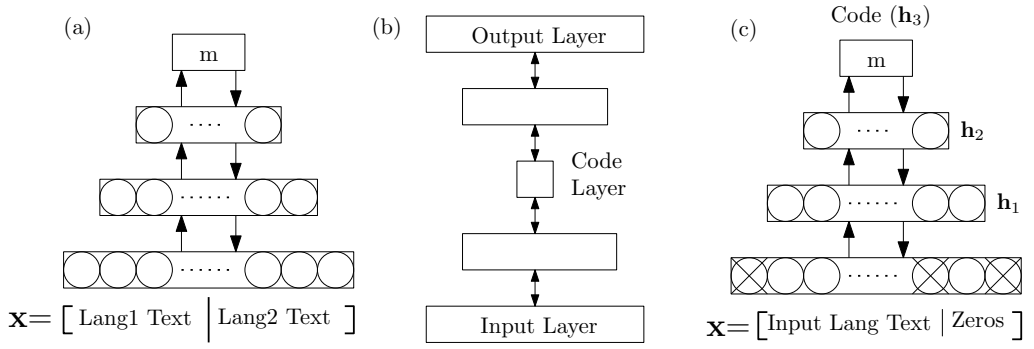

Figure 1: The architecture of the autoencoder during (a) pre-training and (b) fine-tuning, where m is the size of the code layer. Post training, the abstract level representation of the input text can be obtained as shown in (c).

As shown in Fig. 1 (c), representation for the input text $\mathbf{x}$ is obtained as shown below:

$$\begin{aligned} \mathbf{h}_1 &= \sigma(W_1 * \mathbf{x} + \mathbf{b}_1) \\ \mathbf{h}_i &= \sigma(W_i * \mathbf{h}_{i-1} + \mathbf{b}_i), j = 2 \text{ and } 3 \end{aligned} \tag{3}$$

where, W_i and $\mathbf{b}_i$ represent the weight and bias parameters of the layer i and σ is a logistic function to provide non-linearity. For details on training, please refer to Gupta et al. (2014).

3.4. Similarity Learning via Siamese Neural Network (S2Net)

Following the general Siamese neural network architecture (Bromley et al., 1993), S2Net trains two identical neural networks concurrently. The S2Net takes in parallel data with binary or real-valued similarity score and updates the model parameters accordingly (Yih et al., 2011). It optimises a dynamic objective function which is directly modelled by using cosine similarity. The projection operation can be described as follows:

$$y_D = W * x_D \tag{4}$$

where, x_D is the input term vector for document D, W is the learnt projection matrix (represented by the model parameters) and y_D is the latent representation of document D. The parameters of the S2Net are tuned according to the details provided in Yih et al. (2011).

3.5. Machine Translation (MT)

We train a phrase-based machine translation system on the training parallel data using the standard state-of-the-art Moses toolkit[1] with default parameters (Koehn et al., 2007). In this case, the query is translated to the language of documents by MT and then the monolingual similarity is calculated using the BM25 measure[2]. Although we consider this system as a baseline, we do not expect cross-language latent approaches to necessarily outperform it, because this system operates in the original vector space in contrast to latent semantic models, which operate in a low dimensional abstract space.

[1] `http://www.statmt.org/moses/`

[2] We tried different retrieval models like TF-IDF and divergence from randomness based; BM25 performed the best but the difference in performance was not statistically significant

4. Approach

Most prior work on learning low-dimensional semantic representations across languages rely completely on parallel data for training the models (Platt et al., 2010; Yih et al., 2011; Gupta et al., 2014). Our proposed framework removes this requirements by exploiting also monolingual data for model training purposes, and as such it can be more easily applied to low-resource languages.

Specifically, we attempt to incorporate external relevance signals such as the pseudo-relevance data or clickthrough data into the learning framework. Such data might not be available cross-lingually and is mostly confined to the monolingual setting, as most of the present search engines do not employ cross-lingual retrieval explicitly. The main idea behind our proposal is that, monolingual models can be intialised from such largely available relevance data and then, with the help of a smaller amount of parallel data, the cross-lingual model can be trained. This property helps to gain more confidence for under-represented terms in parallel data, *i.e.* terms with very low frequency.

4.1. Monolingual Pre-initialisation

Our proposed learning framework first trains a monolingual model using external relevance data, where the model is encouraged to generate similar representations for relevant documents as measured by cosine similarity. This pre-initialisation can be conducted by using any monolingual latent semantic model from the literature. We consider a model similar to the deep semantic structured model (DSSM) (Huang et al., 2013) with two modifications: *i)* we do not use word-hashing as we will extend this model to the cross-lingual framework and we are more interested in bilingual word associations, and *ii)* while they use a standard bag-of-word vector representation to feed text into the model, we use a composition function.

Consider a function $f : x \rightarrow y \in \mathbb{R}^d$, which embeds a document vector x in vector space to y in d dimensional latent space. We use a simple additive vector composition function on top of the deep neural network output. The architecture of the composition model with m layers is shown in Fig. 2. The input layer accepts the document vector x and the output layer (l_m) provides the semantic representation for the input term vectors. In our approach, we represent each term x_i of the document vector x as *one-hot* representation. Such one-hot vector has the same size as the vocabulary, and only one dimension is on (non-zero). The hidden layer activities and

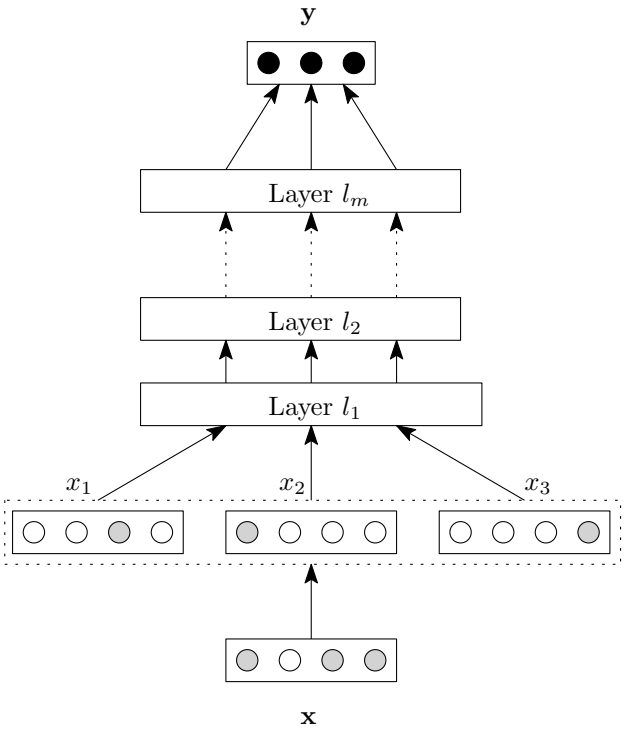

Figure 2: Composition Model. Input text is represented as **x** which as x_1, x_2 and x_3 terms, y is distributed representation of x.

the semantic representation y are obtained as shown in Eq. 5. As it can be noticed in Eq. 5, we perform an additive composition over the representation of terms in the output layer (l_m).

$$\begin{aligned} y_i^{l_1} &= g(W_1 * x_i + b_1) \\ y_i^{l_j} &= g(W_j * y_i^{l_{j-1}} + b_j), j = 2, ..., m \\ y &= \sum_{i=1}^{n} y_i^{l_m} \end{aligned} \tag{5}$$

where, W_j and b_j are j^{th} layer weights and biases respectively, n is the total number of terms in the document and $g(z)$ is a non-linear activation function. In our approach we use the hyperbolic tangent for non-linearity as follows:

$$g(z) = \tanh(z) = \frac{1 - e^{-2z}}{1 + e^{-2z}} \tag{6}$$

This composition framework is slightly different from the standard bag-of-words representation of documents with a feed-forward neural network, because the terms are added after applying the non-linearity which allows to learn word representations directly.

The architecture of the proposed monolingual pre-initialisation model is depicted in Fig. 3. This model is trained to maximise the following objective function,

$$J(\theta) = \cos(y_Q, y_{D^+}) - \cos(y_Q, y_{D^-}) \tag{7}$$

where, $\cos(y_Q, y_D)$ denotes the cosine similarity between the semantic representations of query (Q) and document (D) as shown below:

$$\text{sim}(y_Q, y_D) = \cos(y_Q, y_D) = \frac{\vec{y_Q}^T \vec{y_D}}{\|\vec{y_Q}\| \|\vec{y_D}\|} \tag{8}$$

Maximising the proposed objective function motivates the cosine similarity between relevant document (positive sample, D^+) and query (Q) to be high and the similarity between irrelevant document (negative sample, D^-) and the query (Q) to be low. The noise-contrastive component ($\cos(y_Q, y_{D^-})$) prevents the model from over-fitting and helps to generalise well. Although one can use actual relevant documents for the query as positive samples, they are very few in numbers (a few thousands), hence in practice approaches exploit clickthrough data to proxy actual relevance. In our approach, as we do not have access to clickthrough data, we consider the most relevant document according to the BM25 scoring as a positive sample for the query and the negative sample is selected randomly from the corpus. Our methodology to draw positive samples is motivated by the assumptions of (Rocchio, 1971) algorithm for pseudo relevance feedback. During the training, model parameters are updated using gradient based methods, whose details are presented in Section 5.1. For brevity and consistency, the details of gradient derivation for the objective function in Eq. 7 are given in Appendix A.1.

4.2. Cross-lingual Extension

The main idea of the proposed learning framework is to achieve a cross-lingual representation in a semi-supervised manner from the perspective of parallel data. Given the monolingual composition model already trained on the pseudo-relevance data, a cross-lingual extension can be trained with the use of parallel data. To achieve this, we first project one side of the

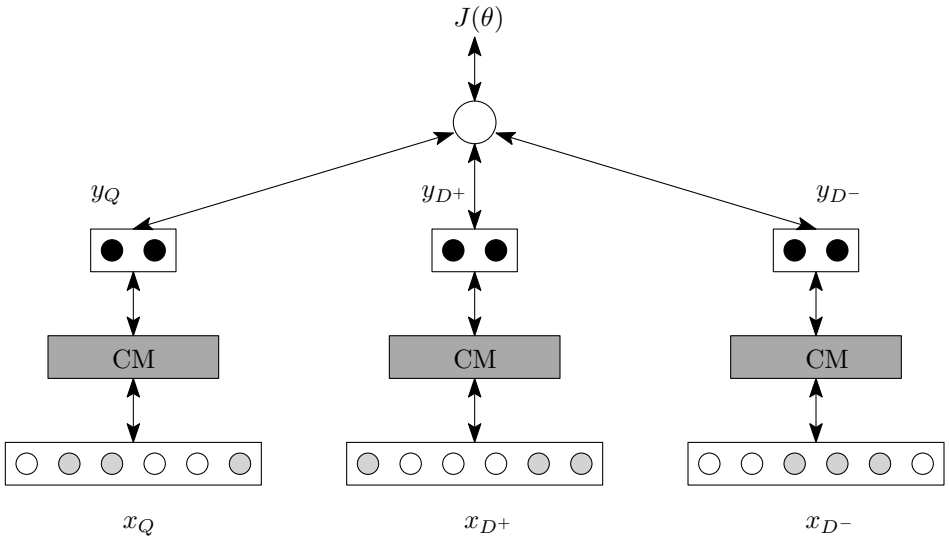

Figure 3: Relevance backpropagation model for monolingual pre-initialisation of the latent space using monolingual relevance data. x_Q, x_{D^+} and x_{D^-} represent input data with corresponding positive and negative samples, while, y_Q, y_{D^+} and y_{D^-} are their distributed representations respectively. It should be noted that there is only one neural network in practice, three instances are displayed for better visualisation.

parallel data by using its corresponding monolingual model. Then, we tune the cross-lingual extension with the use of the other parallel half.

Consider a 3-tuple $(y_{l_1}, y^+_{l_2}, y^-_{l_2})$, where l_1 is the language for which we are training the cross-lingual extension, y_{l_1} denotes the embedding of term vector x in l_1. On the other hand, $y^+_{l_2}$ denotes the embedding of the parallel counterpart of x in l_2 and $y^-_{l_2}$ is the noise component in l_2. Again, the negative sample is chosen randomly from the corpus so that it is irrelevant to the x with maximum probability as done in monolingual pre-initialisation[3]. The architecture of the model is depicted in Fig. 4 and the corresponding objective function is:

$$J_{cl}(\theta) = \cos(y_{l_1}, y^+_{l_2}) - \cos(y_{l_1}, y^-_{l_2}) \tag{9}$$

The composition model CM_{l_2} is obtained through monolingual pre-initialisation.

[3]It is possible for a negative sample to be actually relevant to the input sentence but according to the principles of probability, it is very difficult for it to be consistently relevant or for a large number of datapoints. In practice, randomly selecting a negative sample works well.

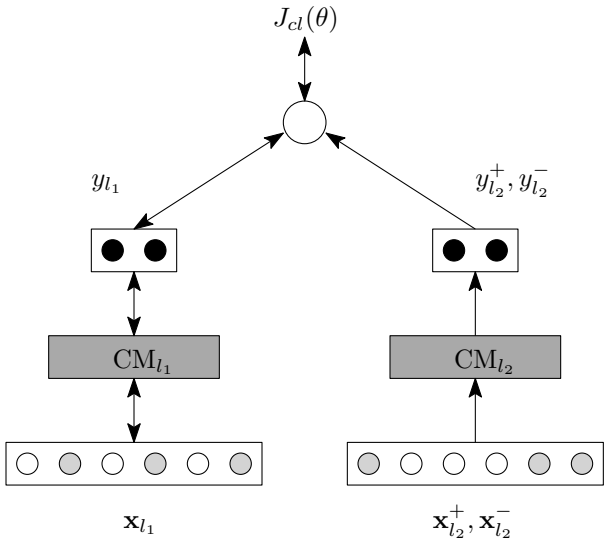

Figure 4: Cross-lingual Extension Model.

It can be noticed from Fig. 4 that only the model parameters of CM_{l_1} are updated during the training. The details of gradient derivation for the objective function presented in Eq. 9 are given in Appendix A.2.

5. Experiments and Results

We evaluated the proposed method, referred as external-data composition neural network (XCNN), and compared it with the existing approaches on two different cross-language tasks for the English-Hindi (En-Hi) language pair: *i)* parallel sentence retrieval, and *ii)* ad-hoc retrieval. First, we explain the experimental setup and training of the proposed model, and then, we give details about the results for the two tasks.

5.1. Learning

For the monolingual pre-initialisation, we use titles from Hindi news articles (∼330k) as queries and get the positive sample for each of them by considering the most relevant title according to the TF-IDF score. These news articles cover different domains (e.g. sports, politics, popular culture, *etc.*) and are collected from Navbharat Times[4]. The cross-lingual systems,

[4]`http://navbharattimes.indiatimes.com/`

all the baselines and our proposed cross-lingual extension, are trained using En-Hi training parallel sentences (∼125k). Details of the parallel corpus are given in Section 5.2.

For the cross-lingual extension, the composition model parameters were initialised randomly under a normal distribution. To keep the model energy low at the beginning, we multiply the parameters by 0.1. During training, we split the data into mini-batches of 100 samples where, each mini-batch can be processed using efficient multi-core CPU/GPU infrastructure. The model parameters are updated after each mini-batch. We use conjugate gradient with 3 iterations and 3 line-searches in each iteration to maximise the objective function. This has been shown to perform well with similar models (Le et al., 2011). For all the latent semantic models, including ours, we consider latent space of 128 dimensions, and raw high dimensional space of 20k dimensions (10k for each language). Although a regularization term can be easily included in Eq. 7 and Eq. 9, we empirically noticed that early stopping was more effective. We did not notice any significant performance difference when more layers were considered, hence $m = 1$ was used. We noticed that monolingual pre-initialisation training converged in roughly 20 epochs and cross-lingual extension converged in around 50 epochs over the entire training data. The GPU based implementation of our proposed model is publicly available at `https://github.com/parthg/jDNN`.

5.2. Parallel Sentence Retrieval

With the advent of the Web, cross-language information retrieval becomes important not only to satisfy the information need across languages but to mine resources for multiple languages, such as parallel or comparable documents. Such mined resources can aid training machine translation systems (Munteanu and Marcu, 2005; Türe and Lin, 2012). The aim of cross-lingual parallel sentence retrieval is to find parallel counterparts into different target languages for a given sentence, or text fragment, in a given source language.

We compare our proposed method with all described baseline systems on the En-Hi parallel corpus available from WMT 2014[5] (Bojar et al., 2014). We extracted the working vocabulary from this corpus by removing stopwords, applying stemming and keeping the most frequent 20k (10k for each

[5]ACL 2014 ninth workshop on statistical machine translation `http://www.statmt.org/wmt14/`.

language). We considered 100k parallel sentences from the corpus, which at least contained 3 terms from the vocabulary, for training and the remaining 21.5k parallel sentences for evaluation. For a fair comparison, all the models were trained and evaluated on the same training and evaluation partitions with the same vocabulary. The results for the sentence retrieval task are presented in Table 1. The retrieval quality for each test sentence is measured by considering its parallel counterpart's reciprocal rank in the ranklist measured by Mean Reciprocal Rank (MRR). This is described in Eq. 10, where Q is the query-set and rank_i is the rank of the first relevant document for query i.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{10}$$

In general, the models with noise-contrastive component outperform the ones without it; e.g. OPCA vs. CL-LSI, and {XCNN, S2Net} vs. BAE. It should also be noted that models such as the S2Net and XCNN, which directly optimise the evaluation metric (cosine similarity) outperform the rest of latent space models such as the CL-LSI, OPCA and BAE. It can be noticed in Table 1, that the proposed method clearly outperforms the other methods with a statistically significant difference (p-value less than 0.01), according to the paired t-test. It should also be noted that the non-linear models outperform the corresponding linear counterparts; e.g. BAE vs. {CL-LSI, OPCA}, and XCNN vs. S2Net.

Method	**MRR**
CL-LSI	0.2620
OPCA	0.4349
BAE	0.4789
S2Net	0.4731
MT	0.4876
XCNN	**0.5328**

Table 1: Results for the parallel sentence retrieval task measured in Mean Reciprocal Rank (MRR).

5.3. Ad-hoc Retrieval

Another task we consider to evaluate the models is the standard ad-hoc retrieval in the cross-language setting. In ad-hoc retrieval, the goal is to

find relevant documents for the user information need specified by the query, typically a few keywords long.

For this evaluation, we tested the models on the standard FIRE 2011-12 En-Hi CLIR track corpus[6]. It contains 100 English queries (topics), 331,599 news articles in Hindi and corresponding relevance judgments (qrels). The retrieval results are evaluated by the standard IR metrics, more specifically, we used mean reciprocal rank (MRR), mean average-precision (MAP) and normalised discounted cumulative gain (nDCG) (Järvelin and Kekäläinen, 2002).

First of all, we checked for the quality of the monolingual pre-initialisation stage, which corresponding results using Hindi queries are presented in Table 2. In the table, BM25 and mono-XCNN are evaluated using a limited vocabulary of size 10k. Interestingly, for the top rank-position related metrics like nDCG@1 and MRR, mono-XCNN performs better than BM25. For other metrics which involve lower rank positions, the performance of mono-XCNN is sub-optimal to the VSM approach. This is not surprising because, in our experimental setting, pseudo relevance data comes from the BM25 scores and mono-XCNN is trained to optimise it; hence it is ought to be upper-bounded by the BM25 scores for lower dimensions. We also expect the gain of XCNN to be higher if clickthrough data is used instead of pseudo relevance data. However, in general this result is also consistent with other works in which it is shown that using only latent models in monolingual setting might hurt the ranking performance, especially for the case of very low dimensional latent space (Manning and Schütze, 1999; Gao et al., 2011).

Method	**nDCG@1**	**nDCG@5**	**nDCG@10**	**MAP**	**MRR**
BM25	0.2800	0.2814	0.2758	0.0957	0.3851
mono-XCNN	0.3000	0.2472	0.2233	0.0794	0.4173

Table 2: Results for the monolingual ad-hoc retrieval task measured in nDCG, MAP and MRR.

For the cross-language setting, the retrieval performance is presented in Table 3 considering the title field of the queries and whole body of the documents. All the models were trained on the training partition of the parallel data described in Sec. 5.2. As computation memory and time scale quadrat-

[6] `http://www.isical.ac.in/~fire/`

ically with the size of vocabulary for models based on eigen decomposition such as the CL-LSI and OPCA, we fixed the cross-language vocabulary size to 20k for all the models. The resulting overall ranking for this task, compared to that on the parallel sentence retrieval task, changes because in this case the relevant documents are not exact translation of the query, but rather represent the main concept of the query. It can be noticed from Table 3 that XCNN outperforms all the models with statistical significance, as measured by a paired t-test (p-value$<$0.05). The linear projection based techniques: CL-LSI, OPCA and S2Net, perform close to each other without significant difference. Also, as seen from the table, the overall results for this task are low. This is mainly because of two reasons: *i)* the selected vocabulary does not cover all the query and document terms, resulting in many out-of-vocabulary (OOV) terms, and *ii)* the parallel training data is not large enough and come from various domains different from that of the FIRE corpus. However, this situation affects equally all the models, which provides a fair ground for comparison. To remove the possible effects due to out-of-vocabulary terms, we recomputed the evaluation metrics considering only those queries for which at least 80% of terms are included in the vocabulary[7]. These results are presented in Table 4.

Method	**nDCG@1**	**nDCG@5**	**nDCG@10**	**MAP**	**MRR**
CL-LSI	0.1200	0.0544	0.0420	0.0062	0.1471
OPCA	0.1300	0.0806	0.0663	0.0254	0.1573
S2Net	0.1263	0.0823	0.0734	0.0278	0.1837
BAE	0.1588	0.1136	0.1057	0.0310	0.2136
MT	0.1800	0.1333	0.1273	**0.0418**	0.2537
XCNN	**0.2200**	**0.1525**	**0.1312**	0.0386	**0.3128**

Table 3: Results for the ad-hoc retrieval task measured in nDCG, MAP and MRR for the title topic field. The best results are highlighted in bold-face.

To the best of our knowledge, this is the first time these latent semantic models are compared on an ah-hoc CLIR task. Usually, S2Net outperforms CL-LSI and OPCA on comparable document retrieval tasks when the S2Net parameters are initialised from the projection matrix of CL-LSI or OPCA, but when it is intialised randomly the gain is smaller (Yih et al., 2011). In

[7]There are 80 such queries out of total 100 queries overall

Method	nDCG@1	nDCG@5	nDCG@10	MAP	MRR
CL-LSI	0.1463	0.0591	0.0416	0.0069	0.1639
OPCA	0.1524	0.0914	0.0762	0.0291	0.1790
S2Net	0.1603	0.1003	0.0826	0.0334	0.2103
BAE	0.1690	0.1129	0.1067	0.0354	0.2332
MT	0.1707	0.1278	0.1224	0.0411	0.2538
XCNN	**0.2683**	**0.1787**	**0.1535**	**0.0459**	**0.3711**

Table 4: Results for the ad-hoc retrieval task measured in nDCG, MAP and MRR for the title topic field considering only those queries for which more than 80% query-terms appear in the vocabulary. The best results are highlighted in bold-face.

this work, we initialised S2Net parameters with weights obtained through OPCA. Though it improves the results for S2Net, as already discussed, computing matrix factorization for CL-LSI and OPCA scale quadratically with vocabulary size, which makes such dependence computationally impractical for high dimensional applications such as the ad-hoc retrieval. Similarly, it is possible to initialise XCNN parameters with the parameters obtained through autoencoders, we wanted to study the abilities of these models to learn semantically plausible representations without dependence on any external method, so we intialised XCNN parameters randomly. Interestingly, our model is also able to outperform MT based method which indicates that our model was able to capture useful cross-lingual semantic representations within a very low dimensional space.

Here we present the implementation level details of our XCNN model. At the time of indexing, each document is represented as a vector in new low-dimensional space using the relevant composition model for that language. It should be noted that a vector for a particular term does not change across documents. Hence, it is efficient to project the vocabulary to the new space once and perform composition for each document to obtain document level representation. At the time of retrieval, a query is also represented in that space and a cosine similarity is calculated between the query and documents. It can be performed as a vector-matrix multiplication. The resulting scores are sorted to present a ranked-list.

Finally, we would like to comment on the efficiency part of the continuous space models for CLIR. The similarity measure for retrieval task, the cosine-distance, is actually semimetric (Skopal and Bustos, 2011). In order to rank documents for the given query, the cosine-distance of all the documents is

calculated *wrt* the query. In vector space, the dimesionality is very high and such distance calculation operation is very expensive. The inverted index provides an efficient way of computing the distance by only considering a subset of documents for which at least one query term is present in the document. The inverted index provides a good example of the suitability of a nonmetric method over metric methods such as Euclidean distance. On the other hand, the continuous space models convert the high dimensional vector space data into continuous low-dimensional data for which we have to scan all the documents linearly. Though the dimensionality is much lower than vector space and therefore, linear scan is still possible, especially with multi-core CPU/GPUs. There are a few efficient alternatives to index such continuous vectors and limit the search space *e.g. IGrid* (Aggarwal and Yu, 2000; Skopal and Bustos, 2011).

6. Conclusions

We have presented and evaluated the external-data composition neural network (XCNN) framework on two different tasks and found the proposed model to be statistically superior in performance to other strong baselines. Especially, the performance is very high for metrics related to top positions - a desired quality for precision oriented systems. The two attributes of the proposed model prove crucial for its performance in the retrieval tasks. First, the learning framework proposed in this work gives a natural way to extend external relevance signals available in the form of pseudo relevance or clickthrough data to cross-language embeddings with the help of a small subset of parallel data. Secondly, the non-linear composition model optimises an objective function that directly relates to the considered task evaluation metric. These properties allow for the model to perform better than other latent semantic models which rely only on parallel data for training.

The gradient based learning provides a way to scale up to large training datasets more easily than linear methods that depend on matrix factorization, such as the CL-LSI and OPCA. For our proposed model, time and space complexity grow linearly with the size of the vocabulary and the amount of training datapoints, while complexity grows quadratically for models based on matrix factorization. Our proposed model also outperforms S2Net, the only latent semantic model that optimises a loss function directly related to the evaluation metric. Although S2Net parameters can be initialised with the projection matrix of CL-LSI or OPCA, such dependence is not practical for

a large vocabulary and large dataset tasks, especially because of the limitations they involve. Moreover, our model can also be initialised with parameters obtained through unsupervised methods, which can potentially improve performance. However, in this work we were more interested in studying the capabilities of the models to learn cross-language embeddings without such dependence. The use of non-linearity allows the model to learn interesting interactions between the terms across languages, within embeddings of dimensionality compared to their linear counterparts. This observation is consistent with the results from other works (Hinton and Salakhutdinov, 2006; Gupta et al., 2014).

Acknowledgements

We thank Germán Sanchis Trilles for helping in conducting experiments with machine translation. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GeForce Titan GPU used for this research. The research of the first author was supported by FPI grant of UPV. The research of the third author is supported by the SomEMBED TIN2015-71147-C2-1-P MINECO research project and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030).

Appendix A. Gradient derivation

In this appendix, we derive the gradient calculation for the model updates. We first show derivation for monolingual pre-initialisation, and then, it is extended for cross-lingual extension.

A.1 monolingual pre-initialisation

The parameters of the monolingual pre-initialisation model are shared among the data points: x_Q, x_{D^+} and x_{D^-} as shown in Fig. 3. As each of them contribute to the objective function in Eq. 7, the gradient can be derived as follows:

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{\partial J(\theta)}{\partial \theta_Q} + \frac{\partial J(\theta)}{\partial \theta_{D^+}} + \frac{\partial J(\theta)}{\partial \theta_{D^-}} \tag{.1}$$

where,

$$\frac{\partial J(\theta)}{\partial \theta_Q} = \frac{\partial \cos(y_Q, y_{D^+})}{\partial \theta_Q} - \frac{\partial \cos(y_Q, y_{D^-})}{\partial \theta_Q} \tag{.2}$$

In the deep neural network architecture, the θ is composed of multiple layer parameters (weights and biases). For example, the gradient of the cosine similarity terms in Eq. .2 at the output layer (L_m) w.r.t. weight matrix W_m with tanh activation can be obtained as follows:

$$\begin{aligned}\frac{\partial \cos(y_Q, y_D)}{\partial \theta_Q^{W_m}} &= \frac{\partial}{\partial \theta_Q^{W_m}} \frac{y_Q^T y_D}{\|y_Q\|\|y_D\|} \\ &= [(1 - y_Q) . * (1 + y_Q) . * \delta_Q^{W_m}] \; y_Q^{L_{m-1}}\end{aligned} \tag{.3}$$

where $.*$ represent element-wise multiplication, and

$$\begin{aligned}\delta_Q^{W_m} &= \frac{1}{\|y_D\|} \frac{\partial}{\partial \theta_Q} \frac{y_Q^T y_D}{\|y_Q\|} \\ &= \frac{1}{\|y_D\|} \left(\frac{\|y_Q\| y_D - (y_Q^T y_D) \frac{y_Q}{\|y_Q\|}}{\|y_Q\|^2} \right) \\ &= \frac{1}{\|y_D\|} \frac{1}{\|y_Q\|} \; y_D - y_Q^T y_D \; \frac{1}{\|y_D\|} \frac{1}{\|y_Q\|^3} \; y_Q\end{aligned} \tag{.4}$$

For clear representation, let scalars $y_Q^T y_D$, $\frac{1}{\|y_Q\|}$ and $\frac{1}{\|y_D\|}$ as a, b and c respectively. Then,

$$\begin{aligned}\frac{\partial \cos(y_Q, y_D)}{\partial \theta_Q^{W_m}} &= [(1 - y_Q) . * (1 + y_Q) . * (bc \; y_D - acb^3 \; y_Q)] \; y_Q^{L_{m-1}} \\ \frac{\partial \cos(y_Q, y_D)}{\partial \theta_D^{W_m}} &= [(1 - y_D) . * (1 + y_D) . * (bc \; y_Q - ac^3 b \; y_D)] \; y_D^{L_{m-1}}\end{aligned} \tag{.5}$$

Putting all together, Eq. .2 becomes:

$$\frac{\partial J(\theta)}{\partial \theta_Q^{W_m}} = [(1 - y_Q) . * (1 + y_Q) . * (bc_p \; y_{D^+} - a_p c_p b^3 \; y_Q - bc_n \; y_{D^-} + a_n c_n b^3 \; y_Q)] \; y_Q^{L_{m-1}} \tag{.6}$$

where $a_p = y_Q^T y_{D^+}$, $c_p = \frac{1}{\|y_{D^+}\|}$, $a_n = y_Q^T y_{D^-}$, $c_n = \frac{1}{\|y_{D^-}\|}$. Similarly for hidden layers, the gradients can be obtained through backpropagation.

A.2 Cross-lingual extension

The parameters of CM_{l_2} are fixed during the cross-lingual extension training, only the parameters of CM_{l_1} contribute to the objective function in Eq. 9 as shown in Fig. 4. Hence, the derivative of the objective function is obtained as follows:

$$\begin{aligned}\frac{\partial J_{cl}(\theta)}{\partial \theta} &= \frac{\partial J_{cl}(\theta)}{\partial \theta_{l_1}} \\ &= \frac{\partial \cos(y_{l_1}, y^+_{l_2})}{\partial \theta_{l_1}} - \frac{\partial \cos(y_{l_1}, y^-_{l_2})}{\partial \theta_{l_1}}\end{aligned} \tag{.7}$$

According to Eq. .6, the gradient at the output layer (L_m) of CM_{l_1} *w.r.t.* W_m can be obtained as follows:

$$\frac{\partial J_{cl}(\theta)}{\partial \theta^{W_m}_{l_1}} = [(1-y_{l_1}).*(1+y_{l_1}).*(bc_p\ y^+_{l_2} - a_p c_p b^3\ y_{l_1} - bc_n\ y^-_{l_2} + a_n c_n b^3\ y_{l_1})]\ y^{L_{m-1}}_{l_1}$$

where $a_p = y^T_{l_1} y^+_{l_2}$, $c_p = \frac{1}{\|y^+_{l_2}\|}$, $a_n = y^T_{l_1} y^-_{l_2}$, $c_n = \frac{1}{\|y^-_{l_2}\|}$.

References

Aggarwal, C. C., Yu, P. S., 2000. The igrid index: Reversing the dimensionality curse for similarity indexing in high dimensional space. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '00. ACM, New York, NY, USA, pp. 119–129.
URL http://doi.acm.org/10.1145/347090.347116

Ballesteros, L., Croft, B., 1996. Dictionary methods for cross-lingual information retrieval. In: Proceedings Of The 7th International Dexa Conference On Database And Expert Systems Applications. pp. 791–801.

Blei, D. M., Ng, A. Y., Jordan, M. I., Mar. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
URL http://dl.acm.org/citation.cfm?id=944919.944937

Bojar, O., Diatka, V., Rychl, P., Stranak, P., Suchomel, V., Tamchyna, A., Zeman, D., may 2014. Hindencorp - hindi-english and hindi-only corpus for machine translation. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland.

Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R., 1993. Signature verification using A "siamese" time delay neural network. IJPRAI 7 (4), 669–688.
URL http://dx.doi.org/10.1142/S0218001493000339

Chandar A. P., S., Lauly, S., Larochelle, H., Khapra, M. M., Ravindran, B., Raykar, V. C., Saha, A., 2014. An autoencoder approach to learning bilingual word representations. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. pp. 1853–1861.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., Harshman, R. A., 1990. Indexing by latent semantic analysis. JASIS 41 (6), 391–407.

Diamantaras, K. I., Kung, S. Y., 1996. Principal Component Neural Networks: Theory and Applications. John Wiley & Sons, Inc., New York, NY, USA.

Dumais, S., Landauer, T. K., Littman, M. L., 1997. Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing. In: AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval. pp. 18–24.

Gabrilovich, E., Markovitch, S., 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference on Artifical intelligence. IJCAI'07. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1606–1611.
URL http://dl.acm.org/citation.cfm?id=1625275.1625535

Gao, J., He, X., Nie, J.-Y., 2010. Clickthrough-based translation models for web search: From word models to phrase models. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. CIKM '10. ACM, New York, NY, USA, pp. 1139–1148.
URL http://doi.acm.org/10.1145/1871437.1871582

Gao, J., Toutanova, K., Yih, W.-t., 2011. Clickthrough-based latent semantic models for web search. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '11. ACM, New York, NY, USA, pp. 675–684.
URL http://doi.acm.org/10.1145/2009916.2010007

Gupta, P., Bali, K., Banchs, R. E., Choudhury, M., Rosso, P., 2014. Query expansion for mixed-script information retrieval. In: The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014. pp. 677–686.

Gupta, P., Costa-Jussà, M. R., Rosso, P., Banchs, R. E., 2016. A deep source-context feature for lexical selection in statistical machine translation. Pattern Recognition Letters 75, 24–29.
URL http://dx.doi.org/10.1016/j.patrec.2016.02.014

Hermann, K. M., Blunsom, P., 2013. The role of syntax in vector space models of compositional semantics. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers. pp. 894–904.

Hiemstra, D., Kraaij, W., Pohlmann, R., Westerveld, T., 2001. Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In: Revised Papers from the Workshop of Cross-Language Evaluation Forum on Cross-Language Information Retrieval and Evaluation. CLEF '00. Springer-Verlag, London, UK, UK, pp. 102–115.
URL http://dl.acm.org/citation.cfm?id=648263.753374

Hinton, G., Salakhutdinov, R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504 – 507.

Hinton, G. E., Salakhutdinov, R. R., 2009. Replicated softmax: an undirected topic model. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. pp. 1607–1614.

Hofmann, T., 1999. Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '99. ACM, New York, NY, USA, pp. 50–57.
URL http://doi.acm.org/10.1145/312624.312649

Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., Heck, L., 2013. Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22Nd ACM International Conference on Conference

on Information & Knowledge Management. CIKM '13. ACM, New York, NY, USA, pp. 2333–2338.
URL http://doi.acm.org/10.1145/2505515.2505665

Järvelin, K., Kekäläinen, J., 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. 20 (4), 422–446.
URL http://doi.acm.org/10.1145/582415.582418

Klementiev, A., Titov, I., Bhattarai, B., 2012. Inducing crosslingual distributed representations of words. In: COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India. pp. 1459–1474.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., 2007. Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. ACL '07. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 177–180.
URL http://dl.acm.org/citation.cfm?id=1557769.1557821

Lauly, S., Boulanger, A., Larochelle, H., 2014. Learning multilingual word representations using a bag-of-words autoencoder. CoRR abs/1401.1803.

Le, Q. V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., Ng, A. Y., 2011. On optimization methods for deep learning. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011. pp. 265–272.

Manning, C. D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013a. Efficient estimation of word representations in vector space. CoRR abs/1301.3781.
URL http://arxiv.org/abs/1301.3781

Mikolov, T., Le, Q. V., Sutskever, I., 2013b. Exploiting similarities among languages for machine translation. CoRR abs/1309.4168.

Mimno, D. M., Wallach, H. M., Naradowsky, J., Smith, D. A., McCallum, A., 2009. Polylingual topic models. In: Proceedings of the 2009 Conference

on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 880–889.

Munteanu, D. S., Marcu, D., Dec. 2005. Improving machine translation performance by exploiting non-parallel corpora. Comput. Linguist. 31 (4), 477–504.
URL http://dx.doi.org/10.1162/089120105775299168

Nie, J.-Y., Simard, M., Isabelle, P., Durand, R., 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '99. ACM, New York, NY, USA, pp. 74–81.
URL http://doi.acm.org/10.1145/312624.312656

Platt, J. C., Toutanova, K., tau Yih, W., 2010. Translingual document representations from discriminative projections. In: EMNLP. pp. 251–261.

Rocchio, J. J., 1971. Relevance feedback in information retrieval. In: Salton, G. (Ed.), The Smart retrieval system - experiments in automatic document processing. Englewood Cliffs, NJ: Prentice-Hall, pp. 313–323.

Salakhutdinov, R., Hinton, G., Jul. 2009. Semantic hashing. Int. J. Approx. Reasoning 50 (7), 969–978.
URL http://dx.doi.org/10.1016/j.ijar.2008.11.006

Skopal, T., Bustos, B., Oct. 2011. On nonmetric similarity search problems in complex domains. ACM Comput. Surv. 43 (4), 34:1–34:50.
URL http://doi.acm.org/10.1145/1978802.1978813

Socher, R., Huval, B., Manning, C. D., Ng, A. Y., 2012. Semantic Compositionality Through Recursive Matrix-Vector Spaces. In: Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Türe, F., Lin, J. J., 2012. Why not grab a free lunch? mining large corpora for parallel sentences to improve translation modeling. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada. pp. 626–630.

Vinokourov, A., Shawe-Taylor, J., Cristianini, N., 2002. Inferring a semantic representation of text via cross-language correlation analysis. In: Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]. pp. 1473–1480.

Xu, J., Croft, W. B., 1996. Query expansion using local and global document analysis. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '96. ACM, New York, NY, USA, pp. 4–11.
URL http://doi.acm.org/10.1145/243199.243202

Yih, W., Toutanova, K., Platt, J. C., Meek, C., 2011. Learning discriminative projections for text similarity measures. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL 2011, Portland, Oregon, USA, June 23-24, 2011. pp. 247–256.

Zou, W. Y., Socher, R., Cer, D. M., Manning, C. D., 2013. Bilingual word embeddings for phrase-based machine translation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL. pp. 1393–1398.