Learning Cross-lingual Representations with Matrix Factorization

Hanan Aldarmaki and **Mona Diab**
Department of Computer Science
The George Washington University
{aldarmaki;mtdiab}@gwu.edu

Abstract

We present a matrix factorization model for learning cross-lingual representations. Using sentence-aligned corpora, the proposed model learns distributed representations by factoring the given data into language-dependent factors and one shared factor. Moreover, the model can quickly learn shared representations for more than two languages without undermining the quality of the monolingual components. The model achieves an accuracy of 88% on English to German cross-lingual document classification, and 0.8 Pearson correlation on Spanish-English cross-lingual semantic textual similarity. While the results do not beat state-of-the-art performance in these tasks, we show that the crosslingual models are at least as good as their monolingual counterparts.

1 Introduction

A large body of NLP research in recent years has focused on representing natural language words and phrases in high-dimensional continuous vector spaces. Such representations can be integrated with various NLP applications as they can be easily learned, processed, and compared, often in an unsupervised or semi-supervised manner. Distributed representations of words, or word embeddings, can be learned using global word co-occurrence statistics as in matrix factorization models (Guo and Diab, 2012; Pennington et al., 2014), or using local context as in neural probabilistic language models (Bengio et al., 2003; Collobert and Weston, 2008; Socher et al., 2013). Compared to word embeddings, representing variable-length sequences using a vector space model is more challenging since these vectors need to encode complex semantic structures and relationships. Several models have been proposed for learning phrase and sentence embeddings, either by combining word embeddings (Klementiev et al., 2012) or directly learning the sentence representations (Le and Mikolov, 2014).

In our global world of information, many NLP problems exist in multilingual and cross-lingual settings. It is often desirable to generalize sentence representations to several languages such that sentences conveying the same meaning in any language are clustered together and potentially mapped to one another in the semantic space. Such cross-lingual representations can then be used directly in NLP applications such as machine translation and cross-lingual question answering. They can also be used to learn classifiers that generalize to languages beyond the ones used in training.

A number of models have recently been proposed for learning cross-lingual compositional representations (Klementiev et al., 2012; Shi et al., 2015; Pennington et al., 2014; Cavallanti et al., 2010; Mikolov et al., 2013; Coulmance et al., 2015; Pham et al., 2015). We propose a relatively simple and nuanced model inspired by the monolingual weighted matrix factorization (WMF) model proposed in (Guo and Diab, 2012), which we extend to the cross-lingual setting.

The WMF model learns word representations by decomposing a sparse tf-idf matrix into two low-rank factor matrices representing words and sen-

Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP, pages 1–9,
San Diego, California, June 17, 2016.

tences. The weights are adjusted to reflect the confidence levels in reconstructing observed vs. missing words in the original matrix. Representations for variable-length sequences can be calculated by minimizing the reconstruction error as described in Section 2.1. In this paper, we propose to extend this model to the cross-lingual setting by modeling two languages in parallel to obtain shared semantic representations. The proposed model has a simple loss function and only uses sentence-aligned data for learning the shared representations. Furthermore, the model can be readily extended to multiple languages without loss of quality. We describe the model in two variations in Section 2.2.

We evaluate the quality of these representations using the cross-lingual document classification task, where a multi-class perceptron is trained to classify documents into four categories. Using German and English labeled short documents, the classifier is trained on one language and tested on the other. Using the compositional representations generated by our model, we achieve an accuracy of 88% in the English→German classification task. We also evaluate on the Semeval cross-lingual semantic textual similarity (STS) task, where we assign a similarity score to pairs of English and Spanish sentences. Our model yields a performance of 0.8 Pearson correlation in this task.

2 Proposed Approach

Word co-occurrence statistics and matrix factorization can be exploited to learn latent semantic representations for words, sentences, and documents (Pennington et al., 2014; Guo and Diab, 2012). We focus on one such model, the weighted matrix factorization model proposed in (Guo and Diab, 2012), as a basis for our crosslingual representations, which is described in the following section. Similar extensions can be implemented for other matrix factorizaiton mehtods.

2.1 Background: Weighted Matrix Factorization (WMF)

In the WMF model proposed in (Guo and Diab, 2012), a large corpus is represented as an $m \times n$ matrix X, where each X_{ij} cell is the tf-idf weight of word i in sentence j. This sparse matrix is then factorized into a $k \times m$ matrix P and a $k \times n$ matrix Q, such that $X = P^T Q$. The factorization results in k-dimensional representations for words and sentences: the columns in P are latent k-dimensional representations for words, and the columns in Q are latent k-dimensional representations for the training sentences.

The values of P and Q can be calculated by minimizing the following weighted loss function:

$$C = \sum_{i,j} W_{ij}(P_{:,i}^T Q_{:,j} - X_{ij})^2 + \lambda(\|P\|^2 + \|Q\|^2) \quad (1)$$

where λ is a regularization parameter to avoid overfitting, and W is an $m \times n$ weight matrix. The weights reflect the confidence levels associated with the reconstruction errors of the corresponding items in X. A small weight is assigned to all missing words, $(X_{ij} = 0)$, to reflect an appropriate level of uncertainty:

$$W_{i,j} = \begin{cases} 1, & \text{if } X_{i,j} \neq 0 \\ w_m, & \text{if } X_{i,j} = 0 \end{cases}$$

where $w_m << 1$ is a fixed weight that is determined empirically; In other words, we assign minimal confidence that each word in the vocabulary could legitimately correlate with any given sentence, while the confidence level is highest for observed words. Using this weighted scheme is explained in more details and experimentally justified in (Guo and Diab, 2012).

By fixing P, the cost function becomes quadratic in Q and the global minimum is achieved using the matrix Q_{min} that satisfies $C'(Q_{min}) = 0$. The j^{th} column in Q_{min} is calculated as follows:

$$Q_{:,j} = (PW^j P^T + \lambda I)^{-1} PW^j X_{:,j} \quad (2)$$

where W^j is a diagonal matrix with coefficients W_{ij} in row/column j (the jth column of W).

Similarly, the vectors in P_{min} are calculated by fixing Q and minimizing the cost function $P(Q)$:

$$P_{:,i} = (QW^i Q^T + \lambda I)^{-1} QW^i X_{i,:} \quad (3)$$

where W^i is a diagonal matrix with coefficients W_{ij} in row/column i (the ith row of W).

Thus, alternating least squares is used to minimize $C(P, Q)$ by iteratively fixing P to calculate Q, then

fixing Q to calculate P using equations (2) and (3). Note that these calculations can be done in parallel and the sparsity of the original matrix can be exploited for a more efficient computation of vectors.[1]

To generate vector representations for additional sentences after training, P is fixed and Q is calculated for the new sentences using equation (2). In other words, we calculate the representations that minimize the loss function (1), which is quadratic when P is fixed.

2.2 Cross-lingual Extensions to WMF

Here we describe our proposed extension of the WMF model for learning bilingual semantic representations. Given a parallel corpus of n sentence pairs, we generate an $m \times n$ tf-idf matrix X for the first language, and an $l \times n$ tf-idf matrix Y for the second language, where m and l are the number of words in the vocabulary of each language. The learning objective of the bilingual WMF model is to factorize both X and Y into two language-specific factors and one shared factor. More precisely, the desired factorization would result in a $k \times m$ matrix P, a $k \times l$ matrix A, and a $k \times n$ matrix Q, such that $X = P^T Q$ and $Y = A^T Q$. To achieve these bilingual objectives, we define two methods for calculating the loss function for both languages as detailed below: A global bilingual loss function (BMF), and a monolingual loss function with an explicit crosslingual factor (CMF).

2.2.1 BMF: Bilingual Matrix Factorization

We define a global loss function for both languages as follows:

$$C=\sum_{i,j} W_{ij}(P_{:,i}^T Q_{:,j}-X_{ij})^2+\sum_{d,j} U_{dj}(A_{:,d}^T Q_{:,j}-Y_{dj})^2 +\lambda(\|P\|^2+\|Q\|^2+\|A\|^2) \quad (4)$$

where U is the weight matrix for Y, defined similarly to W.

This objective function is convex if we fix two of the factor matrices and minimize with respect to the remaining factor. Alternating least squares can be used to estimate the factors iteratively using the following three equations:

$$\begin{aligned} Q_{:,j}&=(PW^j P^T+AU^j A^T+\lambda I)^{-1}(PW^j X_j+AU^j Y_{:,j}) \\ P_{:,i}&=(QW^i Q^T+\lambda I)^{-1} QW^i X_{i,:} \\ A_{:,d}&=(QU^d Q^T+\lambda I)^{-1} QU^d Y_{d,:} \end{aligned} \quad (5)$$

To generate vector representations for additional sentences in either language, the language-specific factors P and A are fixed, and the semantic vectors $Q_{:,j}$ are calculated using equation (2) for language 1 and equation (6) for language 2.

$$Q_{:,j} = (AU^j A^T + \lambda I)^{-1} AU^j Y_{:,j} \quad (6)$$

In other words, the two models are independent once the training is complete, but the resultant representations are expected to reflect shared semantic components.

2.2.2 CMF: Crosslingual Matrix Factorization

Alternatively, we can define two loss functions with a shared crosslingual factor:

$$\begin{aligned} C_1 &= \sum_{i,j} W_{ij}(P_{:,i}^T Q_{:,j} - X_{ij})^2 + \lambda(\|P\|^2 + \|Q\|^2) \\ C_2 &= \sum_{d,j} U_{dj}(A_{:,d}^T Q_{:,j} - Y_{dj})^2 + \lambda(\|A\|^2 + \|Q\|^2) \end{aligned} \quad (7)$$

Minimizing C_1 and C_2 separately is equivalent to training two separate monolingual models. To achieve the bilingual objective, we train only C_1 as a monolingual model, then use the learned factors P to find A. If we assume that the compositional representations generated by P are optimal, then we can use it to fix Q in C_2, and the loss function becomes quadratic in A; all we have to do is find the values of A that minimize C_2.

The training procedure is carried out as follows:

1. Independently train a monolingual WMF model for a pivot language.
2. Using a parallel corpus and the trained word representations P for the pivot language, generate sentence representations Q using equation (2)

[1] Details on similar calculations and speedup recommendations are found in (Hu et al., 2008).

3. Using the same parallel corpus, and fixing Q as calculated in step 2, calculate word representations A for the second language using equation (5).

This method can be readily extended to more than two languages. Using one trained monolingual model, we can quickly learn representations for any number of languages using sentence-aligned data.

3 Related Work

The weighted matrix factorization model we extend was first proposed in (Guo and Diab, 2012) to learn distributed vector representations for words in the monolingual setting. These vectors are then used to generate distributed representations for variable-length sequences by minimizing the reconstruction error. The GloVe algorithm proposed (Pennington et al., 2014) is also a weighted matrix factorization method, but it includes additional word-specific bias terms and uses a different weighting scheme.

As mentioned above, we extend the WMF model proposed in (Guo and Diab, 2012) to bilingual and multilingual settings by forcing the two monolingual components to use a shared factor. (Shi et al., 2015) proposes a similar approach for learning bilingual embeddings. They extend GloVe (Pennington et al., 2014) to the bilingual case using a matrix of bilingual co-occurrence counts with word alignments in addition to the monolingual components. They also propose an alternative method of minimizing the Euclidean distance between words that may be translations of one another. This model is similar in spirit to our model, but it has a different objective function that incorporates cross-lingual co-occurrence statistics or word alignments. Extending that model to more than two languages has not been studied.

In general, several models have been proposed recently to learn cross-lingual semantic representations. Most proposed models learn cross-lingual word embeddings and use them to compose representations for variable-length sequences. For example, (Klementiev et al., 2012) uses a multi-task learning objective (Cavallanti et al., 2010) to align word embeddings for multiple languages. Sentence representations are then composed using idf-weighted sum of word representations. In (Mikolov et al., 2013), word embeddings are first learned separately for each language, and then a linear mapping is learned between the source and target languages.

The `Trans-gram` model introduced in (Coulmance et al., 2015) learns shared representations of several languages efficiently using English as a pivot language. This is comparable to our model in speed and flexibility in learning several language representations. However, the `Trans-gram` model only learns word embeddings, and sentence representations are calculated using idf-weighted average of word embeddings. On the other hand, the WMF model generates sentence representations by minimizing the data reconstruction error in addition to the word embeddings that can be used to calculate an idf-weighted average.

In (Pham et al., 2015), distributed representations for bilingual phrases and sentences are learned using an extended version of the paragraph vector model described in (Le and Mikolov, 2014) by forcing parallel sentences to share one vector. This model learns shared sentence representations directly and achieves state-of-the-art performance in the document classification task.

4 Empirical Evaluation

We evaluate our crosslingual models in two empirical evaluation settings: Crosslingual Semantic Textual Similarity (STS), and Cross-lingual Document Classification (CLDC).

4.1 Data

Monolingual Data: For the monolingual English model, the training set consists of 700K sentences derived from various resources. We extract and combine the following sets: a random set of 150K sentences from LDC's English Gigaword fifth edition (Parker et al., 2011), a random set of 150K sentences from the English Wikipedia[2], the Brown Corpus (Francis, 1964), Wordnet (Miller, 1995) and Wiktionary[3] definitions appended with examples.

Bilingual Data: We extract training data for the bilingual models from WMT13 (Macháček and Bojar, 2013) sentence-aligned parallel corpora, specifically version 7 of the EuroParl parallel corpus (Koehn, 2005), the multiUN parallel corpus (Eisele

[2]http://en.wikipedia.org

[3]http://www.wiktionary.org

and Chen, 2010), and news commentary data for two language pairs: English-Spanish (en-es) and English-German (en-de). We train each bilingual model using a sample of 1M sentence pairs from these datasets.

All sentences in our data are tokenized and stemmed, and number sequences are replaced with a special token as a normalization step. We use the Stanford CoreNLP toolkit (Manning et al., 2014) for English preprocessing, and Treetagger tools (Schmid, 1995) for both Spanish and German data. Words that appear less than 5 times in the training set are discarded from the vocabulary. The final vocabulary sizes for each model are shown below:

Monolingual English	55,881
English from the en-es set	25,057
English from the en-de set	21,879
Spanish	30,411
German	57,431

4.2 Parameter settings for empirical tasks

We train our bilingual BMF models strictly using the bilingual parallel data. On the other hand, we train the English pivot model used in CMF strictly using the English monolingual data, while the parallel corpora are only used for training the Spanish and German components of the CMF models. For the BMF models and the English monolingual model, we run the alternating least squares (ALS) algorithm for 20 iterations. We use the following parameters for all models: k=300, $w_m = 0.01$ and $\lambda = 20$.[4]

4.3 Using English as a Pivot: Cross-lingual STS Validation

One of the advantages of the CMF model is that it can be readily used to learn representations for several languages. We test this hypothesis by using English as a pivot language to learn cross-lingual correspondences between German and Spanish. Hence in this setting, we only use the English model to factor both the Spanish and German models independently, but we assume that any two learned models are directly comparable. Accordingly, WMT12 (Callison-Burch et al., 2012) news test set is used as a validation set to verify that the models actually map the cross-lingual sentences into a shared semantic space. Table 1 shows the average cosine similarity between parallel pairs in the validation set, and the average similarity between a random permutation of the set.

Dataset	en-es	en-de	de-es
Parallel	0.65	0.60	*0.62*
Random	0.11	0.11	0.10

Table 1: Average semantic similarity between sentence pairs using WMT12 test set

These results indicate that the models learn to distinguish between similar and dissimilar sentences since the parallel sentences have much higher cosine similarity than random pairs. We also observe an equivalent performance for the Spanish-German sentences, even though we do not directly train a model for this language pair.

4.4 Cross-lingual Semantic Textual Similarity

Semantic Textual Similarity (STS) is a measure of the degree of similarity between two sentences. STS scores range from 0 to 5, where higher values indicate closer semantic content. Cross-lingual STS measures the degree of similarity between sentences from two different languages.

Using the BMF and CMF cross-lingual models, we generate sentence vectors for the given pairs, then we calculate the cosine similarity between each pair. Since most of the output is positive, and negative values are generally very close to zero, we round up negative similarity values to 0. We then convert the values from the [0-1] range to the [0-5] range by multiplying the scores by 5 [5].

Table 2 shows the results on the test data of Semeval 2016 en-es cross-lingual STS shared task. The evaluation metric is the Pearson Correlation Coefficient. The CMF models perform better than BMF in this task. We also show the results of the official Semeval first rank system, UWB.

4.5 Monolingual Evaluation

We evaluate the performance of the monolingual components learned using BMF or CMF models on the Semeval monolingual Spanish semantic textual similarity (STS) task, namely STS 2014 and STS

[4] These parameters are tuned empirically and we found these values to be robust across models.

[5] Note that this scaling operation does not affect the evaluation results, but we do it for consistency.

Model	News	Multi Source	Mean
BMF	0.83	0.72	0.78
CMF	0.87	0.73	0.80
UWB	0.91	0.82	0.86

Table 2: Cross-lingual STS EN-SP Test results using Pearson Correlation Coefficient.

2015. The objective of this evaluation is to check whether the quality is hurt by forcing the factors into a shared semantic space. We train two monolingual models:

Mono WMF: We train a monolingual Spanish WMF model using the Spanish component of the parallel training set, which consists of 1M sentences. This is the same set used to train the cross-lingual models, so the results are comparable.

WMF*: We train another Spanish WMF model with a more varied training set, similar in construction to the English monolingual model. This training set includes Wikipedia and newswire articles, so it's more similar to the test set. This set consists of about 400K sentences extracted from the second edition of Spanish Gigawords (Mendona et al., 2009) and the Spanish Wikipedia Corpus (Reese et al., 2010).

We use the same values for all the parameters, and we run ALS for 20 iterations. Table 3 shows the results on Semeval Spanish STS 2014 dataset (Agirre et al., 2014), which includes sentence pairs extracted from Spanish Wikipedia and news articles. We also show the results on the harder 2015 dataset (Agirre et al., 2015), which intentionally includes sentence pairs with higher degree of difficulty, such as sentences with shared vocabulary but different compositional meaning. The first row depicts the results obtained by the top system participating in the Semeval task, Semeval Best.

While none of our models outperforms the official Semeval top ranking system, Semeval Best, we show that the Spanish models trained using the BMF and CMF models actually outperform the monolingual Spanish model (mono-WMF) when we use the same dataset for training. The advantage of a monolingual model, however, is that it can be trained using more

Model	WK14	NW14	WK15	NW15
Semeval Best	0.78	0.82	0.71	0.68
WMF*	0.77	0.83	0.64	0.55
mono WMF	0.67	0.80	0.46	0.55
BMF	0.69	0.80	0.50	0.51
CMF	0.70	0.83	0.53	0.52

Table 3: Performance on STS 2014 (WK14, NW14) and STS 2015 (WK15, NW15) test sets for monolingual Spanish STS task

Model	Vector size	en→de	de→en
Maj-Class	40	46.8	46.8
Multi-task	40	77.6	71.1
CLSim	40	92.7	80.2
Trans-gram	100	87.8	78.7
Trans-gram	300	91.1	78.4
BMF	300	88.6	68.9
CMF	300	88.2	70.7
para-doc	500	92.7	91.5

Table 4: Cross-lingual document classification accuracy

varied training data, as evident by the higher performance of `WMF*`, outperforming our cross-lingual derived models.

4.6 Cross-lingual Document Classification (CLDC)

The cross-lingual document classification (CLDC) task introduced in (Klementiev et al., 2012) is a supervised task used to evaluate cross-lingual representations in short document classification. The training and test sets are news stories extracted from the English and German sections of the Reuters multilingual corpus (Lewis et al., 2004). The documents are classified into four categories/topics: C (Corporate/Industrial), E (Economics), G (Government/Social), and M (Markets). For each language, a set of 1K documents is used to train a multi-class Perceptron classifier, and a set of 5K documents is used to test the classifier. For the purpose of evaluating the cross-lingual representations, the classifier is trained on one language and tested on

	to en	to de
from en	88.5	88.2
from de	93.7	70.7

Table 5: Document classification accuracy for CMF model

the other. Thus, we evaluate our models in the English to German direction (en→de), where the model is trained to classify English documents and tested by classifying German documents, and vice versa (de→en).

We generate document representations directly by concatenating all the sentences in each document and using the BMF and CMF models to generate 300-dimensional vectors for each document. The results are shown in Table 4. We show the results of the original `Majority-class` and `Multitask` baselines as listed in (Klementiev et al., 2012). Furthermore, we show the results of several competitive systems on the CLDC task, namely: the `Trans-gram` model (Coulmance et al., 2015), the cross-lingual matrix co-factorization `CLSim` model proposed in (Shi et al., 2015), and the state-of-the art performance by `para-doc` as described in (Pham et al., 2015). We also report the size of the document vector representations for each model.

We note that the performance on the en→de significantly outperforms the other direction of de→en. This trend is apparent in all other models except for `para-doc`. This asymmetry in performance is likely a result of the bag-of-words approach which doesn't account for word ordering. The performance in both directions is lower than that of the competing models, especially in the de→en direction. Also, as shown in table 5, the performance in the crosslingual en→de setting is at least as good as the perforamance in the monolingual en→en setting, while the performance of the de→en crosslingual setting is much lower than the monolingual de→de setting. This indicates that some of the dimensions are not transferred from the German to the English vectors, possible due to unmatched vocabulary cause by the multitude of compound words in German.

5 Discussion and Conclusions

We proposed a new approach for generating crosslingual semantic representations for variable-length sequences using weighted matrix factorization models. These models generally achieved good results in the cross-lingual document classification and crosslingual semantic similarity tasks. One limiting characteristic of the proposed models is the need to use sentence-aligned data, which could undermine the performance in textual domains that lack parallel resources. This can be remedied to some extent by using more representative data in training the pivot model.

A valuable feature of the proposed model is the possibility to learn shared representations for an unlimited number of languages as long as we have sentence-aligned data with one of the learned languages. Training additional languages is trivial since the additional factors are calculated deterministically and independently. In other words, we can learn representations for each language separately and without the need to retrain the available models. Moreover, the model is simple and robust as we learned good representations using relatively small parallel datasets and without parameter optimization. In addition, the monolingual components of the cross-lingual models are as good as, if not better than, the monolingual models learned independently using the same training data. These results direct our attention to the monolingual models we started with; the performance of the crosslingual models is simply a reflection of the quality of the monolingual models they are based on. We focus on improving the monlingual weighted matrix factorization model in future work.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado, June. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and

Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June. Association for Computational Linguistics.

Giovanni Cavallanti, Nicol Cesa-Bianchi, and Claudio Gentile. 2010. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11:2901–2934.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA. ACM.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast crosslingual word embeddings. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1113.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In Daniel Tapias, Mike Rosner, Stelios Piperidis, Jan Odjik, Joseph Mariani, Bente Maegaard, Khalid Choukri, and Nicoletta Calzolari (Conference Chair), editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872. European Language Resources Association (ELRA), 5.

W. Nelson Francis. 1964. A standard sample of present-day english for use with digital computers. *Report to the U.S. Office of Education on Cooperative Research Project No. E-007*.

Weiwei Guo and Mona Diab. 2012. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 864–872, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pages 263–272, Washington, DC, USA. IEEE Computer Society.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196. JMLR Workshop and Conference Proceedings.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, December.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria, August. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Angelo Mendona, David Graff, and Denise DiPersio. 2009. Spanish gigaword second edition ldc2009t21. *Web download file*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition ldc2011t07. *Web download file*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1532–1543.

Hieu Pham, Thang Luong, and Christopher Manning. 2015. Learning distributed representations for multilingual text sequences. *NAACL*.

Samuel Reese, Gemma Boleda Torrent, Montserrat Cuadros Oller, Lluís Padró, and German Rigau Claramunt. 2010. Word-sense disambiguated multilingual wikipedia corpus. In *7th International Conference on Language Resources and Evaluation*.

Helmut Schmid. 1995. Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational*

Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 567–572, Beijing, China, July. Association for Computational Linguistics.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing With Compositional Vector Grammars. In *ACL*.