# GWU_NLP at SemEval-2016 Shared Task 1: Matrix Factorization for Crosslingual STS

**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 for sentences. Using sentence-aligned corpora, the proposed model learns distributed representations by factoring the given data into language-dependent factors and one shared factor. As a result, input sentences from both languages can be mapped into fixed-length vectors and then compared directly using the cosine similarity measure, which achieves 0.8 Pearson correlation on Spanish-English semantic textual similarity.

## 1 Introduction

Semantic textual similarity (STS) is a measure of relatedness in meaning between a pair of variable-length textual snippets, such as sentences. Using unsupervised vector space models, words and sentences can be mapped into dense vector representations that capture implicit syntactic and semantic information. These representations can then be directly compared using well-known distance or similairty measures, such as the Euclidean distance or cosine similarity, which reflect their overall semantic relatedness.

Such 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). A variable-length sentence can be mapped into a fixed-length vector either by combining word embeddings or directly learning sentence representations as in the paragraph vector model proposed in (Le and Mikolov, 2014).

In crosslingual STS, the challenge is to compare sentences from two different languages. We address this problem by directly learning crosslingual vector representations for words and sentences, which allows us to calculate the STS scores without the need for explicit translation or mapping. Several models can be used for learning cross-lingual compositional representations (Klementieva 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 unsupervised 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 sentences. 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. We describe the model in two variations in Section 2.2. This model yields a performance of 0.8 Pearson correlation in



*Proceedings of SemEval-2016*, pages 1247–1251,
San Diego, California, June 16-17, 2016. 

Semeval's English-Spanish crosslingual STS task.

## 2 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 space. 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.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.

## 3 Proposed Approach

### 3.1 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:[1]

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

*where* $\lambda$ 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, $\{X_{ij} \in X | X_{ij} = 0\}$, to reflect some 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* $w_m << 1$; In other words, we assign minimal confidence that each word in the vocabulary could legitimately appear in any given sentence, while the confidence level is highest for observed words.

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 $j$th column of $W$).

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

$$P_i = (QW^i Q^T + \lambda I)^{-1} QW^i {X^T}_i \quad (3)$$

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

Thus, alternating least squares (ALS) 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).[2]

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.

### 3.2 Cross-lingual 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 language 1 sentences, and an $l \times n$ tf-idf matrix $Y$ for

[1] Single subscripts refer to column vectors in all equations.

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

language 2 sentences, 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 (b-WMF), and a monolingual loss function with an explicit shared factor (x-WMF).

#### 3.2.1 b-WMF: Bilingual Loss Function

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 similar 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:

$$\begin{aligned} Q_j &= (PW^j P^T + AU^j A^T + \lambda I)^{-1}(PW^j X_j + AU^j Y_j) \\ P_i &= (QW^i Q^T + \lambda I)^{-1} QW^i X^T{}_i \\ A_d &= (QU^d Q^T + \lambda I)^{-1} QU^d Y^T{}_d \end{aligned} \quad (5)$$

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.

#### 3.2.2 x-WMF: Monolingual Loss Functions

Alternatively, we can define two loss functions with a shared 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 only train $C_1$ as a monolingual model, and then we 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 $C_2$. Given a parallel corpus and word representations $P$, we calculate $Q$ using equation (2), then calculate $A$ using equation (5).

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)

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).

Note that we only use the alternating least squares (ALS) algorithm for training the pivot model; the parameters of the second model, $A$, are calculated deterministically in one step. 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.

## 4 Empirical Evaluation

### 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[3], the Brown Corpus (Francis, 1964), Wordnet (Miller, 1995) and Wiktionary[4] 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 and Chen, 2010), and news commentary data. We train the 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 Spanish. Words that appear less than 5 times in the training set are discarded from the vocabulary.

### 4.2 Parameter Settings

We train our bilingual b-WMF models strictly using the bilingual parallel data. On the other hand, we train the English pivot model used in x-WMF strictly using the English monolingual data, while the parallel corpora are only used for training the Spanish component of the x-WMF models. For the b-WMF models and the English monolingual model, we run the ALS algorithm for 20 iterations. We use the following parameters for all models: $k$=100, $w_m = 0.01$ and $\lambda = 20$.[5]

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

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

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

| Model | News | Multi Source | Mean |
|---|---|---|---|
| b-WMF | 0.83 | 0.72 | 0.78 |
| x-WMF | 0.87 | 0.73 | 0.80 |
| UWB | 0.91 | 0.82 | 0.86 |

**Table 1:** Cross-lingual STS EN-SP Test results using Pearson Correlation Coefficient.

### 4.3 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. In the crosslingual STS test sets, the sentences can either be English or Spanish. We use the TextCat (Cavnar and Trenkle, 1994) tool to identify the languages before processing.

Using the b-WMF and x-WMF 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.

Table 1 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 x-WMF models performs slightly better than b-WMF in this task, and we achieve rank 4 in the official STS semeval evaluation. We also show the results of the official first rank system UWB.

## 5 Discussion and Conclusions

We proposed a new unsupervised approach for generating cross-lingual semantic representations for variable-length sequences using a weighted matrix factorization model. The models successfully learned cross-lingual compositional representations as evident in the high correlation scores in the crosslingual STS task.

Learning a monolingual WMF model involves optimizing a non-convex loss function, but the loss function becomes quadratic once we fix one of the factors. As a result, learning any additional languages becomes trivial once we fix the sentence representations using a pivot model. The model naturally extends to several languages since the additional factors are calculated deterministically. Moreover, the model is simple and robust as we learned

good representations using relatively small parallel datasets and without parameter optimization.

## References


Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March.

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

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

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 Klementieva, Ivan Titov, , and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. *COLING*.

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.

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.

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. *Wen 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*.

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*.