MULTITASK LEARNING OF MULTILINGUAL SENTENCE REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a novel multi-task training approach to learning multilingual distributed representations of text. Our system learns word and sentence embeddings jointly by training a multilingual skip-gram model together with a cross-lingual sentence similarity model. We construct sentence embeddings by processing word embeddings with an LSTM and by taking an average of the outputs. Our architecture can transparently use both monolingual and sentence aligned bilingual corpora to learn multilingual embeddings, thus covering a vocabulary significantly larger than the vocabulary of the bilingual corpora alone. Our model shows competitive performance in a standard cross-lingual document classification task. We also show the effectiveness of our method in a low-resource scenario.

1 INTRODUCTION

Learning distributed representations of text, whether it be at the level of words Mikolov et al. (2013); Pennington et al. (2014), phrases Socher et al. (2010); Pham et al. (2015b), sentences Kiros et al. (2015) or documents Le & Mikolov (2014), has been one of the most widely researched subjects in natural language processing in recent years. Word/sentence/document embeddings, as they are now commonly referred to, have quickly become essential ingredients of larger and more complex NLP systems Bengio et al. (2003); Maas et al. (2011); Collobert et al. (2011); Bahdanau et al. (2014); Chen & Manning (2014) looking to leverage the rich semantic and linguistic information present in distributed representations.

One of the exciting avenues of research that has been taking place in the context of distributed text representations, which is also the subject of this paper, is learning multilingual text representations shared across languages Faruqui & Dyer (2014); Bengio & Corrado (2015); Luong et al. (2015). Multilingual embeddings open up the possibility of transferring knowledge across languages and building complex NLP systems even for languages with limited amount of supervised resources Ammar et al. (2016); Johnson et al. (2016). By far the most popular approach to learning multilingual embeddings is to train a multilingual word embedding model that is then used to derive representations for sentences and documents by composition Hermann & Blunsom (2014). These models are typically trained solely on word or sentence aligned corpora and the composition models are usually simple predefined functions like averages over word embeddings Lauly et al. (2014); Hermann & Blunsom (2014); Mogadala & Rettinger (2016) or parametric coposition models learned along with the word embeddings.

In this work we learn word and sentence embeddings jointly by training a multilingual skip-gram model Luong et al. (2015) together with a cross-lingual sentence similarity model. The multilingual skip-gram model transparently consumes (word, context word) pairs constructed from monolingual as well as sentence aligned bilingual corpora. We use a parametric composition model to construct sentence embeddings from word embeddings. We process word embeddings with a Bi-directional LSTM and then take an average of the LSTM outputs, which can be viewed as context dependent word embeddings. Since our multilingual skip-gram and cross-lingual sentence similarity models are trained jointly, they can inform each other through the shared word embedding layer and promote the compositionality of learned word embeddings at training time. Further, the gradients flowing back from the sentence similarity model can affect the embeddings learned for words outside the vocabulary of the parallel corpora. We hypothesize these two aspects of our model lead to more robust sentence embeddings.

Our contributions are as follows :

- **Scalable approach:** We show that our approach performs better as more languages are added, since represent the extended lexicon in a suitable manner.
- **Ability to perform well in low-resource scenario:** Our approach produces representations comparable with the state-of-art multilingual sentence embeddings using a limited amount of parallel data. Our sentence embedding model is trained end-to-end on a vocabulary significantly larger than the vocabulary of the parallel corpora used for learning cross-lingual sentence similarity.
- **Amenable to Multi-task modeling:** Our model can be trained jointly with proxy tasks, such as sentiment classification, to produce more robust embeddings for downstream tasks.

2 Related Work

This section gives a brief survey of relevant literature. For a through survey of cross-lingual text embedding models, please refer to Ruder (2017).

Cross-lingual Word Embeddings : Most approaches fall into one of these four categories: 1. monolingual mapping: learning transformations from other langauges to English Faruqui & Dyer (2014); Xing et al. (2015); Barone (2016), 2. pseudo cross-lingual: making a pseudo cross-lingual model and training off-the-shelf word embedding models Xiao & Guo (2014); Duong et al. (2016); Vulić & Moens (2016), 3. cross-lingual: learning embeddings using parallel corpora Hermann & Blunsom (2013); Chandar et al. (2014); Søgaard et al. (2015) and 4. joint optimization: using both parallel and monolingual corpora Klementiev et al. (2012); Luong et al. (2015); Vyas & Carpuat (2016); Coulmance et al. (2016). We adopt the skip-gram architecture of Luong et al. (2015) and train a single multilingual model using monolingual data from each language as well as any sentence aligned bilingual data available for any language pair.

Cross-lingual Sentence Embeddings: Some works dealing with cross-lingual word embeddings have considered the problem of constructing sentence embeddings including Vulic & Moens (2015); Pham et al. (2015a); Hermann & Blunsom (2014). In general, it is not trivial to construct cross-lingual sentence embeddings by composing word embeddings as the semantics of a sentence is a complex language-dependent function of its component words as well as their ordering. Pham et al. (2015a) addresses this difficulty by extending the paragraph vector model of Le & Mikolov (2014) to the bilingual context which models the sentence embedding as a separate context vector used for predicting the n-grams from both sides of the parallel sentence pair. At test time, the sentence vector is randomly initialized and trained as part of an otherwise fixed model to predict the n-grams of the given sentence. Our sentence embedding model is closer to the approach taken in Hermann & Blunsom (2014). They construct sentence embeddings by taking average of word or bi-gram embeddings and use a noise-contrastive loss based on euclidean distance between parallel sentence embeddings to learn these embeddings.

Multi-task Learning: Multi-task learning has been employed in various NLP applications where the parameters are shared among tasks Collobert & Weston (2008); Liu et al. (2016); Hashimoto et al. (2016). Liu et al. (2016) show the effectiveness of multi-task learning in multiple sentiment classification tasks by sharing an RNN layer across tasks while learning separate prediction layers for each task. Wu et al. (2017) recently showed benefits of learning a common semantic space for multiple tasks which share a low level feature dictionary. Our multi-task architecture treats training multilingual word embeddings as a separate task with a separate objective as opposed to training them beforehand or training them only as part of a larger model.

3 Model

Our model is trained to optimize two separate objectives: multilingual skip-gram Luong et al. (2015) and cross-lingual sentence similarity. These two tasks are trained jointly with a shared word embedding layer in an end-to-end fashion.

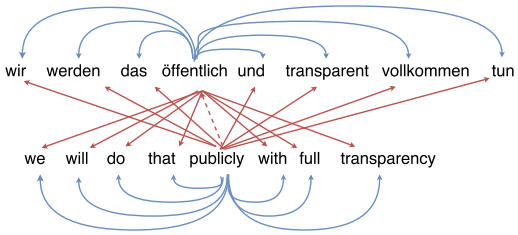

Figure 1: Example context attachments for a bilingual (en-de) skip-gram model.

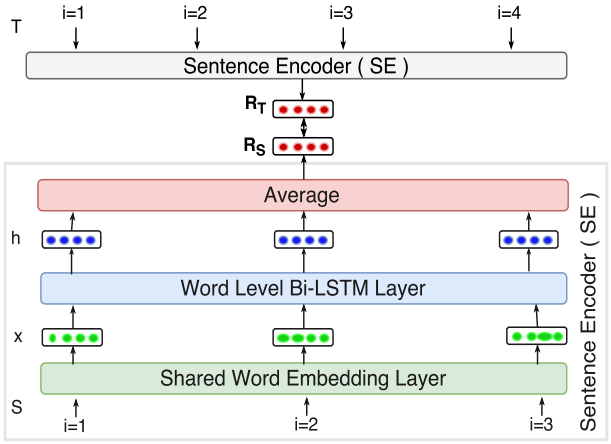

Figure 2: Overview of the architecture that we use for computing sentence representations R_S and R_T for input word sequences S and T.

3.1 Multilingual Skip-gram

Multilingual skip-gram model Luong et al. (2015) extends the traditional skip-gram model by predicting words from both the monolingual and the cross-lingual context. The monolingual context consists of words neighboring a given word as in the case of the traditional skip-gram model. The cross-lingual context, on the other hand, consists of words neighboring the target word aligned with a given source word in a parallel sentence pair. Figure 1, shows an example alignment, where an aligned pair of words are attached to both their monolingual and bilingual contexts. For a pair of languages $L1$ and $L2$, the word embeddings are learned by optimizing the traditional skip-gram objective with (word, context word) pairs sampled from monolingual neighbors in $L1 \rightarrow L1$ and $L2 \rightarrow L2$ directions as well as cross-lingual neighbors in $L1 \rightarrow L2$ and $L2 \rightarrow L1$ directions. In our setup, cross-lingual pairs are sampled from parallel corpora while monolingual pairs are sampled from both parallel and monolingual corpora.

3.2 Cross-lingual Sentence Encoder

We use a parametric composition model to construct sentence embeddings from word embeddings. We process word embeddings with a bi-directional LSTM Hochreiter et al. (2001); Hochreiter & Schmidhuber (1997) and then take an average of the LSTM outputs. There are various implementations of LSTMs available; in this work we use an implementation based on Zaremba et al. (2014). The LSTM outputs (hidden states) contextualize input word embeddings by encoding the history of each word into its representation. We hypothesize that this is better than averaging word embeddings as sentences generally have complex semantic structure and two sentences with different meanings can have exactly the same words. In Figure 2, the word embeddings x_i are processed with a bi-directional LSTM layer to produce h_i. Bi-directional LSTM outputs are then averaged to get a sentence representation.

Learning Method: Let $R : S \rightarrow \mathbb{R}_d$ denote our sentence encoder mapping a given sequence of words S to a continuous vector in $\mathbb{R}_d$. Given a pair of parallel sentences (S, T), we define the loss L of our cross-lingual sentence encoder model as:

$$L_{ST} = \|R_s - R_t\|^2 \tag{1}$$

Therefore, for similar sentences ($S \approx T$), we minimize the loss L_{ST} between their embeddings. We also use a noise-constrastive large-margin update to ensure that the representations of non-aligned sentences observe a certain margin from each other. For every parallel sentence pair (S, T) we randomly sample k negative sentences $N_i, i = 1 \ldots k$. With high probability N_i is not semantically equivalent to S or T.

We define our loss for a parallel sentence pair as follows:

$$\sum_{i=1}^{k} \max(0, m + L_{ST} - L_{SN_i}) \tag{2}$$

Without the LSTM layer, this sentence encoder is similar to the BiCVM Hermann & Blunsom (2014) except that we use also the reversed sample (T, S) to train the model, therefore showing each pair of sentences to the model two times per epoch.

4 CORPORA

Following the literature, we use The Europarl corpus v71 Koehn (2005) for initial development and testing of our approach. We use the first 500K parallel sentences for each of the English-German (en-de), English-Spanish (en-es) and English-French (en-fr) language pairs. We keep the first 90% for training and the remaining 10% for development purposes. We also use additional 500K monolingual sentences from the Europarl corpus for each language. These sentences do not overlap with the sentences in parallel data.

Words which have a frequency less than 5 for a language are replaced with the <unk> symbol. In the joint multi-task setting, the word frequencies are counted using the combined monolingual and parallel corpora. When using just the parallel data for the en-de pair, the vocabulary sizes are 39K for German (de) and 21K for English (en). Vocabulary sizes are 120K for German and 68K for English when both the parallel and the monolingual data are used.

We evaluate our model on the RCV1/RCV2 cross-lingual document classification task where for each language we use 1K documents for training and 5K documents for testing.

5 TRAINING

5.1 TRAINING PARAMETERS

A. Multilingual Skip-gram: We use stochastic gradient descent with a learning rate of 0.01 and exponential decay of 0.98 after 10k steps (1 step = 256 word pairs), negative sampling with 128 samples, skip-gram context window of size 5. Reducing the learning rate of the skip-gram model helps in the multi-task scenario by allowing skip-gram objective to converge in parallel with the sentence similarity objective. We do this modification to make sure that shared word embeddings receive enough supervision from the multilingual sentence similarity objective. At every step, we sample equal number of monolingual and cross-lingual word pairs to make a mini-batch.

B. Sentence Encoder: We keep the batch size to be 50 sentence pairs. LSTM hidden dimension P is one of $100, 128, 512$ depending on the model. We use dropout at the embedding layer with drop probability 0.3. Hinge-loss margin m is always kept to be sentence embedding size. We sample 5 negative samples for the noise-contrastive loss. The model is trained using the Adam optimizer with a learning rate of 0.001 and an exponential decay of 0.98 after 10k steps (1 step = 50 sentence pairs = 1 mini-batch).

The system is optimized by alternating between mini-batches of these two tasks.

5.2 TRAINING ROUTINES

All of our models project words from all input languages to a shared vector space. We train four types of models.

- **Sent-Avg**: This model simply averages word embeddings to get a sentence embedding. It is similar to BiCVM-add model from Hermann & Blunsom (2014), but we also add sentence pairs in the opposite direction, so that the model performs well in both directions.
- **Sent-LSTM**: Represents words in context using the bidirectional LSTM layer, which are then averaged to get sentence embeddings.
- **JMT-Sent-Avg**: Multilingual skip-gram jointly trained with Sent-add. In this setting, the model is optimized by alternating between mini-batches for the two models. JMT refers to Joint Multi-task.
- **JMT-Sent-LSTM**: Multilingual skip-gram jointly trained with Sent-LSTM.

6 EXPERIMENTS

We report results on the Reuters RCV1/RCV2 cross-lingual document classification (CLDC) task Klementiev et al. (2012) using the same experimental setup. We learn the distributed representations on the Europarl corpus.

We construct document embeddings by averaging sentence embeddings. Sentence representations are fixed vectors determined by a sentence encoder trained on parallel and monolingual Europarl corpora. For a language pair $L1$-$L2$, a document classifier (single layer average perceptron) is trained using the document representations from $L1$, and tested on documents from $L2$. Due to lack of supervision on the test side, CLDC setup relies on documents with similar meaning having similar representations.

Table 1, shows the results for our systems and compares it to some state-of-the-art approaches. When the sentence embedding dimension is 128, we outperform most of the systems compared. When the sentence embedding dimension is increased to 512, our results are close to the best results obtained for this task. Our models with an LSTM layer (Sent-LSTM and JMT-Sent-LSTM) are significantly better than those without one. There are also significant gains when the document embeddings are obtained from sentence encoders trained in the multi-task setting. The ablation experiments where we just use parallel corpora suggest that these gains are mostly due to additional monolingual data that we can exploit in the multi-task setting.

Model	en $\rightarrow$ de	de $\rightarrow$ en
dim=128		
BiCVM-ADD	86.4	74.7
BiCVM-BI	86.1	79.0
BiSkip-UnsupAlign	88.9	77.4
Sent-Avg	88.2	80.0
JMT-Sent-Avg	88.5	80.5
Sent-LSTM	**89.5**	80.4
JMT-Sent-LSTM	89.0	**82.2**
JMT-Sent-Avg*no-mono	88.8	80.3
JMT-Sent-LSTM*no-mono	89.0	81.5
dim=500		
para_doc	**92.7**	**91.5**
BiSkip-UnsupAlign	90.7	80.0
Sent-Avg	91.6	84.8
JMT-Sent-Avg	90.8	83.1
Sent-LSTM	92.0	87.3
JMT-Sent-LSTM	92.3	86.2

Table 1: Results for models trained on en-de language pair. *no-mono means no monolingual data was used in training. Dim column gives the dimension of the sentence embeddings. We compare our model to: BiCVM-add+ Hermann & Blunsom (2014), BiCVM-bi+ Hermann & Blunsom (2014), BiSkip-UnsupAlign Luong et al. (2015) and para_doc Pham et al. (2015a).

Model	en-es	en-de	de-en	es-en	es-de
Sent-Avg-(en,es,de,fr)	49.8	86.8	78.4	63.5	69.4
Sent-LSTM-(en,es,de,fr)	53.1	**89.9**	77.0	67.8	65.3
JMT-Sent-Avg-(en,es,de,fr)	51.5	87.2	75.7	60.3	**72.6**
JMT-Sent-LSTM-(en,es,de,fr)	**56.4**	89.7	75.1	63.3	68.1
JMT-Sent-LSTM	53.1	89.0	**82.2**	**68.4**	-

Table 2: We compare our JMT-Sent-LSTM model trained on three languages to one trained on two languages.

6.1 Single Model for multiple languages

Table 2 compares models trained on data from four languages (en, es, de, fr) to models trained on data from two languages. The results suggest that models trained on multiple languages perform better when English is the source language used to train the CLDC system. The multilingual systems also show promising results for es-de pair, for which there was no direct parallel data available.

6.2 Low-resource scenario

Model	Dim	en-de	de-en
Sent-add	128	81.6	75.2
JMT-Sent-add	128	85.3	79.1
Sent-LSTM	128	82.1	76.0
JMT-Sent-LSTM	128	87.4	80.7

Table 3: JMT models show big gains when comparing Sent-add and Sent-LSTM trained on just 100k parallel sentences.

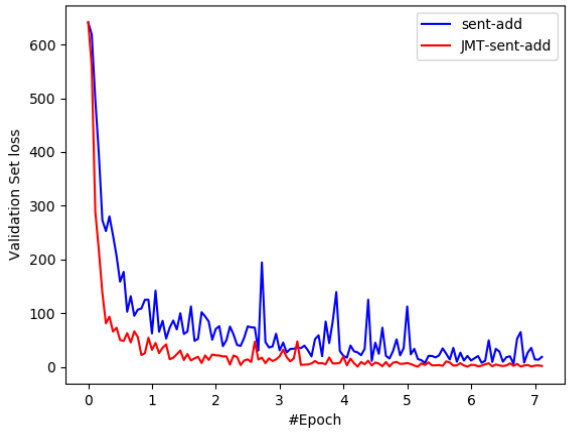

Figure 3: Validation loss for JMT-sent-add model shows more stability and achieves a lower value than the one for Sent-add model in the low-resource scenario. At every training step, the validation set is created by randomly choosing 50 sentences from the development set.

The main motivation behind the multi-task architecture is to create high quality multilingual embeddings for languages which have limited amount of parallel data available. Therefore, we compare the effectiveness of our Joint multi-task models in the low resource scenario, where for each language pair we use 100k parallel sentences and 1 million monolingual sentences for training the sentence encoder. We evaluate on the RCV1/RCV2 document classification task. Like before, we keep the first 90% (90k parallel sentences) of parallel data for training and 10% (10k parallel sentences) for development purposes.

Table 3 shows that JMT training for 128 dimensional sentence embeddings gives big gains in terms of RCV1/RCV2 document classification accuracy when we use only 100k parallel sentences. JMT-Sent-LSTM model results are similar to (en-de) or better than (de-en) the results reported for the BICVM model, which uses 500k parallel sentences for training. These results suggest that JMT model can produce high quality multilingual embeddings without large amounts of parallel data. We believe that this gain is due to the extra monolingual data, hence larger vocabulary, that the JMT model can use transparently.

Figure 3 shows the loss curves for sent-add and JMT-Sent-add models. On the validation set, JMT-Sent-add model gives a smoother and lower loss curve.

7 Discussion and Future Work

Our results suggest that using a parametric composition model to derive sentence embeddings from word embeddings and joint multi-task learning of multilingual word and sentence embeddings are promising directions. This paper is a snapshot of our current efforts and w e believe that our sentence embedding models can be improved further with straightforward modifications to the model architecture, for instance by using stacked LSTMs, and we plan to explore these directions in future work.

In our exploration of architectures for the sentence encoding model, we also tried using a self-attention layer following the intuition that not all words are equally important for the meaning of a sentence. However, we later realized that the cross lingual sentence similarity objective is at odds with what we want the attention layer to learn. When we used self attention instead of simple averaging of word embeddings, the attention layer learns to give the entire weight to a single word in both the source and the target language since that makes optimizing cross lingual sentence similarity objective easier.

Even though they are related tasks, multilingual skip-gram and cross-lingual sentence similarity models are always in a conflict to modify the shared word embeddings according to their objectives. This conflict, to some extent, can be eased by careful choice of hyper-parameters. This dependency on hyper-parameters suggests that better hyper-parameters can lead to better results in the multi-task learning scenario. We have not yet tried a full sweep of the hyperparameters of our current models but we believe there may be easy gains to be had from such a sweep especially in the multi-task learning scenario.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. Many languages, one parser. *arXiv preprint arXiv:1602.01595*, 2016.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Antonio Valerio Miceli Barone. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. *arXiv preprint arXiv:1608.02996*, 2016.

Yoshua Bengio and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. 2015.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pp. 1853–1861, 2014.

Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *EMNLP*, pp. 740–750, 2014.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160–167. ACM, 2008.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. Trans-gram, fast cross-lingual word-embeddings. *arXiv preprint arXiv:1601.02502*, 2016.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Learning crosslingual word embeddings without bilingual corpora. *arXiv preprint arXiv:1606.09403*, 2016.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. Association for Computational Linguistics, 2014.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*, 2016.

Karl Moritz Hermann and Phil Blunsom. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*, 2013.

Karl Moritz Hermann and Phil Blunsom. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*, 2014.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*, 2016.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pp. 3294–3302, 2015.

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

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pp. 79–86, 2005.

Stanislas Lauly, Alex Boulanger, and Hugo Larochelle. Learning multilingual word representations using a bag-of-words autoencoder. *arXiv preprint arXiv:1401.1803*, 2014.

Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pp. 1188–1196, 2014.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*, 2016.

Thang Luong, Hieu Pham, and Christopher D Manning. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 151–159, 2015.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 142–150. Association for Computational Linguistics, 2011.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Aditya Mogadala and Achim Rettinger. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of NAACL-HLT*, pp. 692–702, 2016.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pp. 1532–1543, 2014.

Hieu Pham, Minh-Thang Luong, and Christopher D Manning. Learning distributed representations for multilingual text sequences. In *Proceedings of NAACL-HLT*, pp. 88–94, 2015a.

Nghia The Pham, Germán Kruszewski, Angeliki Lazaridou, and Marco Baroni. Jointly optimizing word representations for lexical and sentential tasks with the c-phrase model. In *ACL (1)*, pp. 971–981, 2015b.

Sebastian Ruder. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902, 2017. URL `http://arxiv.org/abs/1706.04902`.

Richard Socher, Christopher D Manning, and Andrew Y Ng. Learning continuous phrase representations and syntactic parsing with recursive neural networks. In *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, pp. 1–9, 2010.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. Inverted indexing for cross-lingual nlp. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*, 2015.

Ivan Vulic and Marie-Francine Moens. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, pp. 719–725. ACL, 2015.

Ivan Vulić and Marie-Francine Moens. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994, 2016.

Yogarshi Vyas and Marine Carpuat. Sparse bilingual word representations for cross-lingual lexical entailment. In *Proceedings of NAACL-HLT*, pp. 1187–1197, 2016.

Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. Starspace: Embed all the things! *arXiv preprint arXiv:1709.03856*, 2017.

Min Xiao and Yuhong Guo. Distributed word representation learning for cross-lingual dependency parsing. In *CoNLL*, pp. 119–129, 2014.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *HLT-NAACL*, pp. 1006–1011, 2015.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.