A Multi-task Approach to Learning Multilingual Representations

Karan Singla[1], **Dogan Can**[1], **Shrikanth Narayanan**[1,2]
[1]Department of Computer Science
[2]Department of Electrical Engineering
University of Southern California, Los Angeles, USA
{singlak, dogancan}@usc.edu, shri@sipi.usc.edu

Abstract

We present a novel multi-task modeling 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. 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 limited resource scenario.

1 Introduction

Learning distributed representations of text, whether it be at the level of words, phrases, sentences or documents has been one of the most widely researched subjects in natural language processing in recent years (Mikolov et al., 2013; Pennington et al., 2014; Gouws et al., 2015; Socher et al., 2010; Pham et al., 2015b; Kiros et al., 2015; Conneau et al., 2017; Le and Mikolov, 2014; Chen, 2017; Wu et al., 2017). Word/sentence/document embeddings, as they are now commonly referred to, have quickly become essential ingredients of larger and more complex NLP systems looking to leverage the rich semantic and linguistic information present in distributed representations (Bengio et al., 2003; Maas et al., 2011; Collobert et al., 2011; Bahdanau et al., 2014; Chen and Manning, 2014).

Research that has been taking place in the context of distributed text representations is learning multilingual text representations shared across languages (Faruqui and Dyer, 2014; Bengio and Corrado, 2015; Luong et al., 2015). Multilingual embeddings open up the possibility of transferring knowledge across languages and building complex 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 and 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 and Blunsom, 2014; Mogadala and Rettinger, 2016) or parametric composition models learned along with the word embeddings (Schwenk et al., 2017). For a thorough survey of cross-lingual text embedding models, please refer to (Ruder, 2017).

In this work we learn word and sentence embeddings jointly by training a multilingual skip-gram model together with a cross-lingual sentence similarity model. Our multilingual skip-gram model is similar to (Luong et al., 2015). It transparently consumes *(word, context)* pairs constructed from monolingual as well as sentence aligned bilingual corpora. We process word embeddings with a bidirectional LSTM and then take an average of the LSTM outputs, which can be viewed as context dependent word embeddings, to produce sentence 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

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 214–220
Melbourne, Australia, July 15 - 20, 2018.

affect the embeddings learned for words outside the vocabulary of the parallel corpora. We hypothesize these two aspects of approach lead to more robust sentence embeddings.

The main motivation behind our approach is to learn high quality multilingual sentence and document embeddings in the low resource scenario where parallel corpus sizes are limited. The main novelty of our approach is the joint training of multilingual skip-gram and cross-lingual sentence similarity objectives with a shared word embedding layer which allows the gradients from the sentence similarity task to affect the embeddings learned for words outside the vocabulary of the parallel corpora. By jointly training these two objectives, we can transparently use monolingual and parallel data for learning multilingual sentence embeddings. Using a BiLSTM layer to contextualize word embeddings prior to averaging is orthogonal to the joint multi-task learning idea. We observed that this additional layer is beneficial in most settings and this is consistent with the observations of recent works on learning sentence and document embeddings such as (Conneau et al., 2017; Yang et al., 2016)

2 Model

Our model jointly optimizes multilingual skip-gram (Luong et al., 2015) and cross-lingual sentence similarity objectives using a shared word embedding layer in an end-to-end fashion.

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.

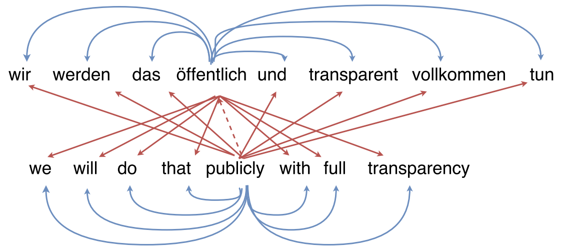

Figure 1: Example context attachments for a bilingual (en-de) skip-gram model.

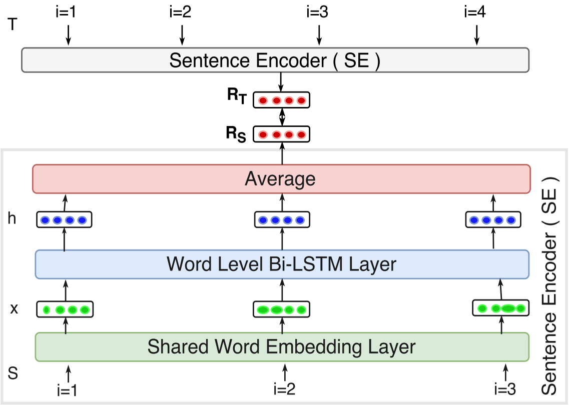

Figure 2: Overview of the architecture that we use for computing sentence representations R_S and R_T for input word sequences S and T.

Cross-lingual Sentence Similarity: We process word embeddings with a bi-directional LSTM (Hochreiter et al., 2001; Hochreiter and Schmidhuber, 1997) and then take an average of the LSTM outputs (Figure 2). 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. 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 their distance as $d(S, T) = \|R_S - R_T\|^2$. For every parallel sentence pair, we randomly sample k negative sentences $\{N_i | i = 1 \dots k\}$ and define the cross-lingual sentence similarity loss as follows:

$$l(S,T) = \sum_{i=1}^{k} \max(0, m + d(S,T) - d(S,N_i))$$

Without the LSTM layer, this loss is similar to the BiCVM loss (Hermann and 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.

3 Experiments

3.1 Corpora

We learn the distributed representations on the Europarl corpus v71 (Koehn, 2005). For a fair comparison with literature, 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 that occur less than 5 times are replaced with the <unk> symbol. In the joint multi-task setting, the words are counted in the combined monolingual and parallel corpora. The vocabulary sizes for German (de) and English (en) are respectively 39K and 21K in the parallel corpus, 120K and 68K in the combined corpus.

We evaluate our models on the RCV1/RCV2 cross-lingual document classification task (Klementiev et al., 2012), where for each language we use 1K documents for training and 5K documents for testing.

3.2 Models

In addition to the proposed joint multi-task (JMT) model, **JMT-Sent-LSTM**, we also present ablation experiments where we omit the LSTM layer, the multilingual skip-gram objective or both. **JMT-Sent-Avg** is like the proposed model but does not include an LSTM layer. **Sent-LSTM** and **Sent-Avg** are the single-task variants of these models.

We construct document embeddings by averaging sentence representations produced by a trained sentence encoder. For a language pair $L1$-$L2$, a document classifier (single layer average perceptron) is trained on documents from $L1$, and tested on documents from $L2$. Due to lack of supervision on the $L2$ side, this setup relies on documents from different languages with similar meaning having similar representations.

3.3 Training

The single-task models are trained with the cross-lingual sentence similarity objective end-to-end using parallel data only. We also tried training word embeddings beforehand on parallel and mono data and tuning them on the cross-lingual sentence similarity task but that did not improve the results. Those results are omitted for brevity. The multi-task models are trained by alternating between the two tasks.

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 is 256 word pairs), negative sampling with 512 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. At every step, we sample equal number of monolingual and cross-lingual word pairs to make a mini-batch.

Cross-lingual Sentence Similarity: The batch size is 50 sentence pairs. LSTM hidden state dimension is 128 or 512. We use dropout at the embedding layer with drop probability 0.3. Hinge-loss margin m is equal to sentence embedding size. We sample 10 negative samples for the noise-contrastive loss. The model is trained using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001 and an exponential decay of 0.98 after 10K steps (1 step is 50 sentence pairs).

3.4 Results

Table 1 shows the results for our models and compares them to some state-of-the-art approaches. When the sentence embedding dimension is 512, our results are close to the best results from literature. When the sentence embedding dimension is 128, our JMT-Sent-LSTM model outperforms all of the systems compared. Models with an LSTM layer (Sent-LSTM and JMT-Sent-LSTM) perform better than those without one. Joint multi-task training consistently improves the performance. The results for the data ablation experiments (*no-mono) suggest that the gains obtained in the JMT setting are partly due to the addition of monolin-

Model	en → de	de → en
500k parallel sentences, dim=128		
BiCVM-add+	86.4	74.7
BiCVM-bi+	86.1	79.0
BiSkip-UnsupAlign	88.9	77.4
Our Models		
Sent-Avg	88.2	80.0
JMT-Sent-Avg	88.5	80.5
Sent-LSTM	89.5	80.4
JMT-Sent-LSTM	**90.4**	**82.2**
JMT-Sent-Avg*no-mono	88.8	80.3
JMT-Sent-LSTM*no-mono	89.5	81.5
100k parallel sentences, dim=128		
Sent-Avg	81.6	75.2
JMT-Sent-Avg	85.3	79.1
Sent-LSTM	82.1	76.0
JMT-Sent-LSTM	87.4	80.7
JMT-Sent-LSTM*no-mono	83.4	76.5

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

gual data and partly due to the multi-task objective.

Varying monolingual vs parallel data: The main motivation behind the multi-task architecture is to create high quality embeddings in the limited resource scenario. The bottom section of Table 1 shows the results for 128 dimensional embeddings when parallel data is limited to 100K sentences. JMT-Sent-LSTM results in this scenario are comparable to the results from the middle section of Table 1 which use 500K parallel sentences. These findings suggest that JMT-Sent-LSTM model can produce high quality embeddings even with a limited amount of parallel data by exploting additional monolingual data. Table 2 compares Sent-LSTM vs. JMT-Sent-LSTM at different data conditions. JMT-Sent-LSTM produces consistently better embeddings as long as the amount of additional monolingual data is neither too large nor too small compared to the amount of parallel data – 3-4 times parallel data size seems to be a good heuristic for choosing monolingual data size.

Multilingual vs Bilingual models: Table 3 compares multilingual models (en, es, de) to bilingual models. First four rows of Table 3 show results for multilingual systems where sentence encoder is trained for three languages (en,es,de) using en-es and en-de parallel data and additional monolingual data for each language. Document representations obtained from this sentence encoder are then used to train a classifier for a language pair like en-de, where the classifier is trained on en documents and then tested on de documents. In this scenario, we can build classifiers for language pairs like es-de even though we do not have access to es-de parallel data since embeddings we learn are shared between the three languages. Bottom row in Table 3 shows results for bilingual systems where we train the sentence encoder for two languages, and then use that encoder to train a document classifier for one language and test on the other. In this scenario, we cannot build classifiers for language pairs like es-de for which we do not have access to parallel data.

Parallel / Mono	20K	50K	100K	500K
no-mono	60.3	68.3	82.1	89.5
20K	57.4	68.7	80.2	89.5
50K	**62.7**	69.0	83.5	89.5
100K	61.5	71.9	85.1	89.6
200K	58.1	**72.1**	85.5	90.0
500K	52.6	64.8	**87.4**	**90.4**

Table 2: Sent-LSTM vs. JMT-Sent-LSTM at different data conditions (en-de, dim=128).

Model	en-es	en-de	de-en	es-en	es-de
Sent-Avg	49.8	86.8	78.4	63.5	69.4
Sent-LSTM	53.1	89.9	77.0	67.8	65.3
JMT-Sent-Avg	51.5	87.2	75.7	60.3	**72.6**
JMT-Sent-LSTM	**57.4**	**91.0**	75.1	63.3	68.1
JMT-Sent-LSTM*	54.1	90.4	**82.2**	**68.4**	-

Table 3: Multilingual vs. bilingual* models (dim=128).

Multilingual models perform better than bilingual ones when English is the source language but they perform worse in the other direction. We believe this discrepancy is because Europarl documents were originally in English and later translated to other languages. The multilingual models also show promising results for es-de pair, for which there was no parallel data.

4 Linguistic analysis

As classification experiments focused on keeping semantic information in sentence level representations, we also checked if produced word embeddings still made sense. We use JMT-Sent-LSTM

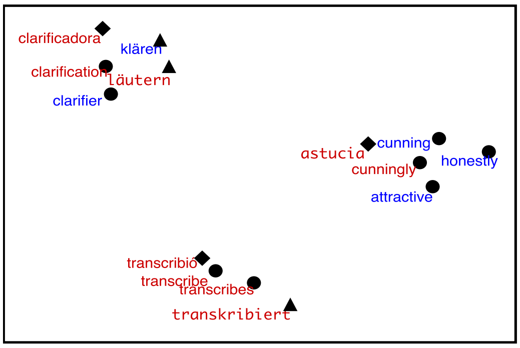

Figure 3: t-SNE projections for 3 English words (clarification, transcribe, cunningly) which are not in the parallel corpus and their four nearest neighbors. Red words are only in the monolingual corpus. Blue words exist in parallel corpus too.

model for this purpose. Figure 3 shows t-SNE projections for some sample words. Even though the model didn't use any German-Spanish parallel data it managed to map words which have similar meaning (transkribiert and transcribi) closer. Words that are antonyms but still have a similar meaning are close to each other (cunnigly (en), honestly (en) and astucia (es)). Nearest neighbors in the multilingual representation space are generally of same form across languages. It can also be observed that English words lie towards the middle of Spanish and German words which we believe is due to English being the pivot for the other two languages.

5 Conclusion

Our results suggest that joint multi-task learning of multilingual word and sentence embeddings is a promising direction. We believe that our sentence embedding model can be improved further with straightforward modifications to the sentence encoder architecture, for instance using stacked LSTMs or batch/layer normalization, and addition of sentence level auxiliary tasks such as sentiment classification or natural language inference. We plan to explore these directions and evaluate our approach on additional tasks in the future.

6 Discussion and 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. Another approach could be to derive high dimensional embeddings in a way similar to (Conneau et al., 2017) and using max-pooling which can allow efficient selection for each dimension to represent meaning.

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 hyper-parameters 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. Other thing that remains rather unexplored is to do other levels of multi-tasking, like learning character representations or multitasking at sentence level.

References

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

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

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

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

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

Minmin Chen. 2017. Efficient vector representation for documents through corruption. *arXiv preprint arXiv:1707.02377*.

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

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

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

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning*, pages 748–756.

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

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

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

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

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 *MT summit*, volume 5, pages 79–86.

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

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

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

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

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

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

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

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

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

Sebastian Ruder. 2017. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902.

Holger Schwenk, Ke Tran, Orhan Firat, and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *arXiv preprint arXiv:1704.04154*.

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

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

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*, pages 1480–1489.

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