Exploring Multilingual Syntactic Sentence Representations

Chen Liu, Anderson de Andrade, Muhammad Osama
Wattpad
Toronto, ON, Canada
cecilia, anderson, muhammad.osama@wattpad.com

Abstract

We study methods for learning sentence embeddings with syntactic structure. We focus on methods of learning syntactic sentence-embeddings by using a multilingual parallel-corpus augmented by Universal Parts-of-Speech tags. We evaluate the quality of the learned embeddings by examining sentence-level nearest neighbours and functional dissimilarity in the embedding space. We also evaluate the ability of the method to learn syntactic sentence-embeddings for low-resource languages and demonstrate strong evidence for transfer learning. Our results show that syntactic sentence-embeddings can be learned while using less training data, fewer model parameters, and resulting in better evaluation metrics than state-of-the-art language models.

1 Introduction

Recent success in language modelling and representation learning have largely focused on learning the semantic structures of language (Devlin et al., 2018). Syntactic information, such as part-of-speech (POS) sequences, is an essential part of language and can be important for tasks such as authorship identification, writing-style analysis, translation, etc. Methods that learn syntactic representations have received relatively less attention, with focus mostly on evaluating the semantic information contained in representations produced by language models.

Multilingual embeddings have been shown to achieve top performance in many downstream tasks (Conneau et al., 2017; Artetxe and Schwenk, 2018). By training over large corpora, these models have shown to generalize to similar but unseen contexts. However, words contain multiple types of information: semantic, syntactic, and morphologic. Therefore, it is possible that syntactically different passages have similar embeddings due to their semantic properties. On tasks like the ones mentioned above, discriminating using patterns that include semantic information may result in poor generalization, specially when datasets are not sufficiently representative.

In this work, we study methods that learn sentence-level embeddings that explicitly capture syntactic information. We focus on variations of sequence-to-sequence models (Sutskever et al., 2014), trained using a multilingual corpus with universal part-of-speech (UPOS) tags for the target languages only. By using target-language UPOS tags in the training process, we are able to learn sentence-level embeddings for source languages that lack UPOS tagging data. This property can be leveraged to learn syntactic embeddings for low-resource languages.

Our main contributions are: to study whether sentence-level syntactic embeddings can be learned efficiently, to evaluate the structure of the learned embedding space, and to explore the potential of learning syntactic embeddings for low-resource languages.

We evaluate the syntactic structure of sentence-level embeddings by performing nearest-neighbour (NN) search in the embedding space. We show that these embeddings exhibit properties that correlate with similarities between UPOS sequences of the original sentences. We also evaluate the embeddings produced by language models such as BERT (Devlin et al., 2018) and show that they contain some syntactic information.

We further explore our method in the few-shot setting for low-resource source languages without large, high quality treebank datasets. We show its transfer-learning capabilities on artificial and real low-resource languages.

Lastly, we show that training on multilingual parallel corpora significantly improves the learned

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 153–159
Hong Kong, Nov 4, 2019. ©2019 Association for Computational Linguistics

syntactic embeddings. This is similar to existing results for models trained (or pre-trained) on multiple languages (Schwenk, 2018; Artetxe and Schwenk, 2018) for downstream tasks (Lample and Conneau, 2019).

2 Related Work

Training semantic embeddings based on multilingual data was studied by MUSE (Conneau et al., 2017) and LASER (Artetxe and Schwenk, 2018) at the word and sentence levels respectively. Multi-task training for disentangling semantic and syntactic information was studied in (Chen et al., 2019). This work also used a nearest neighbour method to evaluate the syntactic properties of models, though their focus was on disentanglement rather than embedding quality.

The syntactic content of language models was studied by examining syntax trees (Hewitt and Manning, 2019), subject-object agreement (Goldberg, 2019), and evaluation on syntactically altered datasets (Linzen et al., 2016; Marvin and Linzen, 2018). These works did not examine multilingual models.

Distant supervision (Fang and Cohn, 2016; Plank and Agic, 2018) has been used to learn POS taggers for low-resource languages using cross-lingual corpora. The goal of these works is to learn word-level POS tags, rather than sentence-level syntactic embeddings. Furthermore, our method does not require explicit POS sequences for the low-resource language, which results in a simpler training process than distant supervision.

3 Method

3.1 Architecture

We iterated upon the model architecture proposed in LASER (Artetxe and Schwenk, 2018). The model consists of a two-layer Bi-directional LSTM (BiLSTM) encoder and a single-layer LSTM decoder. The encoder is language agnostic as no language context is provided as input. In contrast to LASER, we use the concatenation of last hidden and cell states of the encoder to initialize the decoder through a linear projection.

At each time-step, the decoder takes an embedding of the previous POS target concatenated with an embedding representing the language context, as well as a max-pooling over encoder outputs. Figure 1 shows the architecture of the proposed model.

Table 1: Hyperparameters

Parameter	Value
Number of encoder layers	2
Encoder forward cell size	128
Encoder backward cell size	128
Number of decoder layers	1
Decoder cell size	512
Input BPE vocab size	40000
BPE embedding size	100
UPOS embedding size	100
Language embedding size	20
Dropout rate	0.2
Learning rate	1e-4
Batch size	32

The input embeddings for the encoder were created using a jointly learned Byte-Pair-Encoding (BPE) vocabulary (Sennrich et al., 2016) for all languages by using sentencepiece[1].

3.2 Training

Training was performed using an aligned parallel corpus. Given a source-target aligned sentence pair (as in machine translation), we:

1. Convert the sentence in the source language into BPE
2. Look up embeddings for BPE as the input to the encoder
3. Convert the sentence in a target language into UPOS tags, in the tagset of the target language.
4. Use the UPOS tags in step 3 as the targets for a cross-entropy loss.

Hence, the task is to predict the UPOS sequence computed from the translated input sentence.

The UPOS targets were obtained using StandfordNLP (Qi et al., 2018) [2]. Dropout with a drop probability of 0.2 was applied to the encoder. The Adam optimizer (Kingma and Ba, 2015) was used with a constant learning rate of 0.0001. Table 1 shows a full list of the hyperparameters used in the training procedure.

3.3 Dataset

To create our training dataset, we followed an approach similar to LASER. The dataset contains 6

[1] https://github.com/google/sentencepiece

[2] https://stanfordnlp.github.io/stanfordnlp/index.html

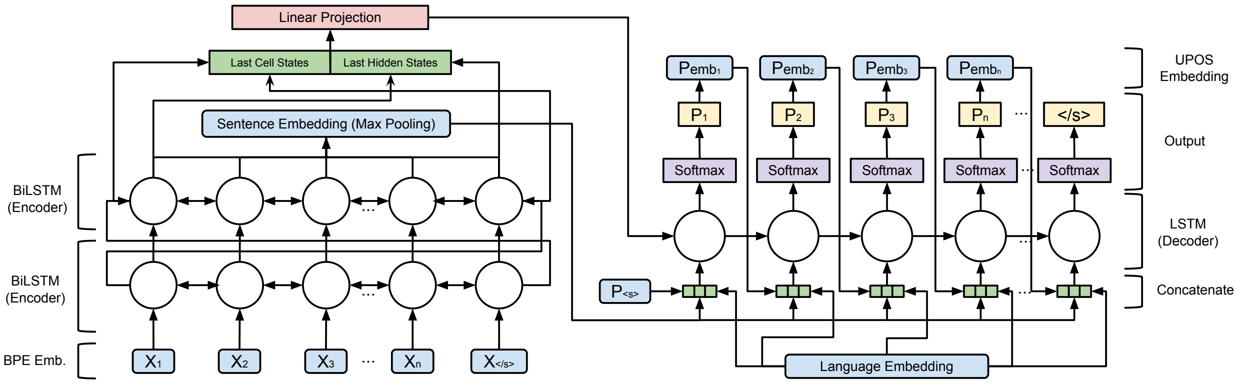

Figure 1: Proposed architecture.

languages: English, Spanish, German, Dutch, Korean and Chinese Mandarin. These languages use 3 different scripts, 2 different language orderings, and belong to 4 language families.

English, Spanish, German, and Dutch use a Latin-based script. However, Spanish is a Romantic language while the others are Germanic languages. Chinese Mandarin and Korean are included because they use non-latin based scripts and originate from language families distinct from the other languages. Although the grammatical rules vary between the selected languages, they share a number of key characteristics such as the *Subject-Verb-Object* ordering, except Korean (which mainly follows the *Subject-Object-Verb* order). We hope to extend our work to other languages with different scripts and sentence structures, such as Arabic, Japanese, Hindi, etc. in the future.

The dataset was created by using translations provided by Tatoeba[3] and OpenSubtitles[4] (Lison and Tiedemann, 2016). They were chosen for their high availability in multiple languages.

Statistics of the final training dataset are shown in Table 2. Rows and columns correspond to source and target languages respectively.

3.3.1 Tatoeba

Tatoeba is a freely available crowd-annotated dataset for language learning. We selected all sentences in English, Spanish, German, Dutch, and Korean. We pruned the dataset to contain only sentences with at least one translation to any of the other languages. The final training set contains 1.36M translation sentence pairs from this source.

3.3.2 OpenSubtitles

We augmented our training data by using the 2018 OpenSubtitles dataset. OpenSubtitles is a publicly available dataset based on movie subtitles (Lison and Tiedemann, 2016). We created our training dataset from selected aligned subtitles by taking the unique translations among the first million sentences, for each aligned parallel corpus. We further processed the data by pruning to remove samples with less than 3 words, multiple sentences, or incomplete sentences. The resulting dataset contains 1.9M translation sentence pairs from this source.

4 Experiments

We aim to address the following questions:

1. Can syntactic structures be embedded? For multiple languages?
2. Can parallel corpora be used to learn syntactic structure for low-resource languages?
3. Does multilingual pre-training improve syntactic embeddings?

We address question 1 in Secs. 4.1 and 4.2 by evaluating the quality of syntactic and semantic embeddings in several ways. Questions 2 and 3 are addressed in Sec. 4.3 by studying the transfer-learning performance of syntactic embeddings.

4.1 Quality of Syntactic Embeddings

We studied the quality of the learned syntactic embeddings by using a nearest-neighbour (NN) method.

First, we calculated the UPOS sequence of all sentences in the Tatoeba dataset by using a tagger. Sentences were then assigned to distinct groups according to their UPOS sequence, i.e., all sentences belonging to the same group had the same

[3] https://tatoeba.org/eng/

[4] http://opus.nlpl.eu/OpenSubtitles-v2018.php

Table 2: Training Dataset Statistics

	English	German	Spanish	Chinese	Korean	Dutch
English	-	521.87k	194.51k	41.33k	31.81k	190.86k
German	520.64k	-	217.96k	5.67k	0.21k	12.20k
Spanish	193.01k	217.46k	-	159.67k	28.68k	144.82k
Chinese	40.79k	5.62k	159.73k	-	0.05k	0.32k
Korean	31.05k	1.37k	28.89k	0.07k	-	56.93k
Dutch	215.18k	25.75k	155.35k	0.66k	56.92k	-

UPOS sequence.

For all languages except Korean, a held-out test set was created by randomly sampling groups that contained at least 6 sentences. For Korean, all groups containing at least 6 sentences were kept as the test set since the dataset is small.

During evaluation, we applied max-pooling to the outputs of the encoder to obtain the syntactic embeddings of the held-out sentences[5].

For each syntactic embedding, we find its top nearest neighbour (1-NN) and top-5 nearest neighbours (5-NN) in the embedding space for the held-out sentences, based on their UPOS group.

Given n sentences $S = \{s_0, \dots, s_{n-1}\}$ and their embeddings $E = \{e_0, \dots, e_{n-1}\}$, for each s_i there is a set of k gold nearest neighbours $G(i,k) = \{g_0, \dots, g_{k-1}\}$, $G(i,k) \subseteq S$ such that $d(s_i, g) \leq d(s_i, s)$ for all $g \in G(i,k)$ and $s \in S \setminus G(i,k)$, where $d(\cdot,\cdot)$ is the cosine distance.

Given embedding e_i, we calculate cosine distances $\{d(e_i, e_j) \text{ for } e_j \in E, e_j \neq e_i\}$ and sort them into non-decreasing order $d_{j_0} \leq d_{j_1} \leq \cdots \leq d_{j_{n-2}}$. We consider the ordering to be unique as the probability of embedding cosine distances being equal is very small.

The set of embedded k-nearest neighbours of s_i is defined as

$$N(i,k) = \{s_j \text{ for } j \in \{j_0, \dots, j_{k-1}\}\}.$$

Finally, the k-nearest neighbours accuracy for s_i is given by

$$\frac{|N(i,k) \cap G(i,k)|}{k}.$$

A good embedding model should cluster the embeddings for similar inputs in the embedding space. Hence, the 5-NN test can be seen as an indicator of how cohesive the embedding space is.

Table 3: Syntactic Nearest-Neighbour Accuracy (%)

	ISO	1-NN/5-NN	Total/Groups
English	en	97.27/93.36	2784/160
German	de	93.45/86.77	1282/91
Spanish	es	93.81/86.24	1503/81
Chinese	zh	71.26/61.44	167/22
Korean	ko	28.27/18.40	527/40
Dutch	nl	74.17/51.71	3171/452

The results are shown in Table 3. The differences in the number of groups in each language are due to different availabilities of sentences and sentence-types in the Tatoeba dataset.

The high nearest neighbours accuracy indicates that syntax information was successfully captured by the embeddings. Table 3 also shows that the syntactic information of multiple languages was captured by a single embedding model.

4.1.1 Language Model

A number of recent works (Hewitt and Manning, 2019; Goldberg, 2019) have probed language models to determine if they contain syntactic information. We applied the same nearest neighbours experiment (with the same test sets) on a number of existing language models: Universal Sentence Encoder (USE) (Cer et al., 2018), LASER, and BERT. For USE we used models available from TensorHub[6]. For LASER we used models and created embeddings from the official repository [7].

For BERT, we report the results using max (BERT$_{max}$) and average-pooling (BERT$_{avg}$), obtained from the BERT embedding toolkit[8] with the multilingual cased model (104 languages, 12-layers, 768-hidden units, 12-heads), and 'pooled-output' (BERT$_{output}$) from the TensorHub version

[5] Evaluation data will be hosted at https://github.com/ccliu2/syn-emb

[6] https://www.tensorflow.org/hub

[7] https://github.com/facebookresearch/LASER

[8] https://github.com/imgarylai/bert-embedding

of the model with the same parameters.

We computed the nearest neighbours experiment for all languages in the training data for the above models. The results are shown in Table 4. The results show that general purpose language models do capture syntax information, which varies greatly across languages and models.

The nearest neighbours accuracy of our syntactic embeddings in Table 3 significantly outperforms the general purpose language models. Arguably these language models were trained using different training data. However, this is a reasonable comparison because many real-world applications rely on released pre-trained language models for syntactically related information. Hence, we want to show that we can use much smaller models trained with direct supervision, to obtain syntactic embeddings with similar or better quality. Nonetheless, the training method used in this work can certainly be extended to architectures similar to BERT or USE.

4.2 Functional Dissimilarity

The experiments in the previous section showed that the proposed syntactic embeddings formed cohesive clusters in the embedding space, based on UPOS sequence similarities. We further studied the spatial relationships within the embeddings.

Word2Vec (Mikolov et al., 2013) examined spatial relationships between embeddings and compared them to the semantic relationships between words. Operations on vectors in the embedding space such as $King - Man + Woman = Queen$ created vectors that also correlated with similar operations in semantics. Such semantic comparisons do not directly translate to syntactic embeddings. However, syntax information shifts with edits on POS sequences. Hence, we examined the spatial relationships between syntactic embeddings by comparing their cosine similarities with the edit distances between UPOS sequence pairs.

Given n UPOS sequences $U = \{u_0, ..., u_{n-1}\}$, we compute the matrix $L \in \mathbb{R}^{n \times n}$, where $l_{ij} = l(u_i, u_j)$, the complement of the normalized Levenshtein distance between u_i and u_j.

Given the set of embedding vectors $\{e_0, ..., e_{n-1}\}$ where e_i is the embedding for sentence s_i, we also compute $D \in \mathbb{R}^{n \times n}$, where $d_{ij} = d(e_i, e_j)$. We further normalize d_{ij} to be within $[0, 1]$ by min-max normalization to obtain $\hat{D} = \text{minMax}(D)$.

Following (Yin and Shen, 2018), we define the *functional dissimilarity score* by

$$\frac{\|L - \hat{D}\|_\text{F}}{n}.$$

Intuitively, UPOS sequences that are similar (smaller edit distance) should be embedded close to each other in the embedding space, and embeddings that are further away should have dissimilar UPOS sequences. Hence, the functional dissimilarity score is low if the relative changes in UPOS sequences are reflected in the embedding space. The score is high if such changes are not reflected.

The functional dissimilarity score was computed using sentences from the test set in CoNLL 2017 Universal Dependencies task (Nivre et al., 2017) for the relevant languages with the provided UPOS sequences. Furthermore, none of the evaluated models, including the proposed method, were trained with CoNLL2017 data.

We compared the functional dissimilarity scores of our syntactic representations against embeddings obtained from BERT and LASER, to further demonstrate that simple network structures with explicit supervision may be sufficient to capture syntactic structure. All the results are shown in Table 5. We only show the best (lowest) results from BERT.

4.3 Transfer Performance of Syntactic Embeddings

Many NLP tasks utilize POS as features, but human annotated POS sequences are difficult and expensive to obtain. Thus, it is important to know if we can learn sentences-level syntactic embeddings for low-sources languages without treebanks.

We performed zero-shot transfer of the syntactic embeddings for French, Portuguese and Indonesian. French and Portuguese are simulated low-resource languages, while Indonesian is a true low-resource language. We reported the 1-NN and 5-NN accuracies for all languages using the same evaluation setting as described in the previous section. The results are shown in Table 6 (top).

We also fine-tuned the learned syntactic embeddings on the low-resource language for a varying number of training data and languages. The results are shown in Table 6 (bottom). In this table, the low-resource language is denoted as the 'source', while the high-resource language(s) is denoted as the 'target'. With this training method, no UPOS

Table 4: Syntactic Nearest-Neighbour for Language Models (%)

	English	German	Spanish	Chinese	Korean	Dutch
Model	1-NN/5-NN	1-NN/5-NN	1-NN/5-NN	1-NN/5-NN	1-NN/5-NN	1-NN/5-NN
USE	71.83/55.68	59.87/44.26	53.05/38.06	39.23/30.18	21.22/12.43	28.66/12.77
BERT$_{max}$	**90.19/86.36**	**83.66/77.63**	**83.89/79.92**	**67.96/68.40**	20.30/11.92	37.67/19.51
BERT$_{avg}$	89.06/84.70	79.54/74.82	78.24/75.61	65.75/67.07	20.30/11.47	37.04/19.46
BERT$_{output}$	77.75/63.44	66.20/51.89	65.21/50.41	52.49/46.34	16.39/10.98	24.27/10.67
LASER	86.33/76.66	76.56/62.88	72.49/59.72	56.89/45.15	**26.63/15.90**	**50.75/31.00**

Table 5: Functional Dissimilarity Scores (Lower is Better)

Model	English	German	Spanish	Chinese	Korean	Dutch
BERT$_{avg}$	0.3463	0.3131	0.2955	0.2935	0.3001	0.3131
LASER	0.1602	0.1654	0.2074	0.3099	0.2829	0.1654
Proposed Work	0.1527	0.1588	0.1588	0.2267	0.2533	0.1588

tag information was provided to the model for the 'source' languages, where supervising information comes solely from parallel sentences and UPOS tags in high-resource languages.

The results show that for a new language (French and Portuguese) that is similar to the family of pre-training languages, there are two ways to achieve higher 1-NN accuracy. If the number of unique sentences in the new language is small, accuracy can be improved by increasing the size of the parallel corpora used to fine-tune. If only one parallel corpus is available, accuracy can be improved by increasing the number of unique sentence-pairs used to fine-tune.

For a new language that is dissimilar to the family of pre-training languages, e.g. Indonesian in Table 6, the above methods only improved nearest neighbours accuracy slightly. This may be caused by differing data distribution or by tagger inaccuracies. The results for Indonesian do indicate that some syntactic structure can be learned by using our method, even for a dissimilar language.

A future direction is to conduct a rigorous analysis of transfer learning between languages from the same versus different language families.

5 Conclusion

We examined the possibility of creating syntactic embeddings by using a multilingual method based on sequence-to-sequence models. In contrast to prior work, our method only requires parallel corpora and UPOS tags in the target language.

We studied the quality of learned embeddings by examining nearest neighbours in the embedding space and investigating their functional dissimilarity. These results were compared against recent state-of-the-art language models. We also showed that pre-training with a parallel corpus allowed the syntactic embeddings to be transferred to low-resource languages via few-shot fine-tuning.

Our evaluations indicated that syntactic structure can be learnt by using simple network architectures and explicit supervision. Future directions include improving the transfer performance for low-resource languages, disentangling semantic and syntactic embeddings, and analyzing the effect of transfer learning between languages belong to the same versus different language families.

Table 6: Syntactic Nearest-Neighbour on New languages (%)

Lang (ISO)	1-NN/5-NN	Total/Group
French (fr)	35.86/22.11	6816/435
Protuguese (pt)	48.29/23.15	4608/922
Indonesian (id)	21.00/35.92	657/59

	Number of Parallel Sentence Pairs	
Source -Target(s)	2k	10k
ISO	1-NN/5-NN	1-NN/5-NN
fr-en	47.37/32.18	58.41/42.87
fr-(en,es)	46.82/31.92	58.01/42.65
pt-en	56.75/30.14	64.52/36.94
pt-(en,es)	57.94/30.63	65.00/37.06
id-en	27.09/47.64	31.35/56.01

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. A multi-task approach for disentangling syntax and semantics in sentence representations. In *NAACL2019*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Meng Fang and Trevor Cohn. 2016. Learning when to trust distant supervision: An application to low-resource POS tagging using cross-lingual projection. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *NAACL2019*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *EMNLP2018*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and James A. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Joakim Nivre, Željko Agić, Lars Ahrenberg, et al. 2017. Universal dependencies 2.0. lindat/clarin digital library at the institute of formal and applied linguistics, charles university, prague.

Barbara Plank and Zeljko Agic. 2018. Distant supervision from disparate sources for low-resource part-of-speech tagging. In *EMNLP2018*.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Zi Yin and Yuanyuan Shen. 2018. On the dimensionality of word embedding. In *Advances in Neural Information Processing Systems*, pages 887–898.