Scalable Cross-Lingual Transfer of Neural Sentence Embeddings

Hanan Aldarmaki[1] and **Mona Diab**[1,2]
[1]The George Washington University
[2]AWS, Amazon AI
aldarmaki@gwu.edu,diabmona@amazon.com

Abstract

We develop and investigate several cross-lingual alignment approaches for neural sentence embedding models, such as the supervised inference classifier, InferSent, and sequential encoder-decoder models. We evaluate three alignment frameworks applied to these models: joint modeling, representation transfer learning, and sentence mapping, using parallel text to guide the alignment. Our results support representation transfer as a scalable approach for modular cross-lingual alignment of neural sentence embeddings, where we observe better performance compared to joint models in intrinsic and extrinsic evaluations, particularly with smaller sets of parallel data.

1 Introduction

Probabilistic sentence representation models generally fall into two categories: bottom-up compositional models, where sentence embeddings are composed from word embeddings via a linear function like averaging, and top-down compositional models that are trained with a sentence-level objective, typically within a neural architecture. Sequential data like sentences can be modeled using recurrent, recursive, or convolutional networks, which can implicitly learn intermediate sentence representations suitable for each learning task. Depending on the training objective, these intermediate representations sometimes encode enough semantic and syntactic features to be suitable as general-purpose sentence embeddings. For examples, it was shown in Conneau et al. (2017a) that a model trained to maximize inference classification accuracy can yield generic representations that perform well across a wide set of extrinsic classification benchmarks. Other training objectives, like denoising auto-encoders or neural sequence to sequence models (Hill et al., 2016), can also yield general-purpose representations with different characteristics. While bottom-up models can achieve superior performance in tasks that are independent of syntax, such as topic categorization, neural models often yield representations that encode syntactic and positional features, which results in superior performance in tasks that rely on sentence structure (Aldarmaki and Diab, 2018).

General-purpose sentence embeddings can be used as features in various classification tasks, or to directly assess the similarity of a pair of sentences using the cosine measure. It is often desired to generalize word and sentence embeddings across several languages to facilitate cross-lingual transfer learning (Zhou et al., 2016) and mining of parallel sentences (Guo et al., 2018). For word embeddings, cross-lingual learning can be achieved in various ways (Upadhyay et al., 2016), such as learning directly with a cross-lingual objective (Shi et al., 2015) or post-hoc alignment of monolingual word embeddings using dictionaries (Ammar et al., 2016), parallel corpora (Gouws et al., 2015; Klementiev et al., 2012), or even with no bilingual supervision (Conneau et al., 2017b; Aldarmaki et al., 2018). For bottom-up composition like vector averaging, word-level alignment is sufficient to yield cross-lingual sentence embeddings. For top-down sentence embeddings, the efforts in cross-lingual learning are more limited. Typically, a multi-faceted cross-lingual learning objective is used to align the sentence models while training them, as in Soyer et al. (2014). Cross-lingual sentence embeddings can also be learned via a neural machine translation framework trained jointly for multiple languages (Schwenk and Douze, 2017).

While they indeed yield cross-lingual embeddings, the joint training models in existing literature pose some practical limitations: simultaneous training requires massive computational re-

sources, particularly for sequential models like the bi-directional LSTM networks typically used to encode sentences. In addition, the joint framework does not allow post-hoc or modular training, where new languages can be added and aligned to existing pre-trained encoders. More recently, Conneau et al. (2018) proposed an approach for cross-lingual sentence embeddings by aligning encoders of new languages to a pre-trained English encoder using parallel corpora. Such approach promises to be more suitable for modular training of general sentence encoders, although so far it has only been evaluated in natural language inference classification.

In this paper, we develop and evaluate three alignment frameworks: joint modeling, representation transfer learning, and sentence mapping, applied on two modern general-purpose sentence embedding models: the inference-based encoder, InferSent (Conneau et al., 2017a), and the sequential denoising auto-encoder, SDAE (Hill et al., 2016). For most approaches, we rely on parallel sentences as sentence-level dictionaries for cross-lingual supervision. We report the performance on sentence translation retrieval and cross-lingual document classification. Our results support representation transfer as a scalable approach for modular cross-lingual alignment that works well across different neural models and evaluation benchmarks.

2 Related Work

Learning bilingual compositional representations can be achieved by optimizing a bilingual objective on parallel corpora. In Pham et al. (2015), distributed representations for bilingual phrases and sentences are learned using an extended version of the paragraph vector model (Le and Mikolov, 2014) by forcing parallel sentences to share one vector. In Soyer et al. (2014), cross-lingual compositional embeddings are learned by optimizing a joint bilingual objective that aligns parallel source and target representations by minimizing the Euclidean distances between them, and a monolingual objective that maximizes the similarity between similar phrases. The monolingual objective was implemented by maximizing the similarity between random phrases and subphrases within the same sentence. Cross-lingual representations can also be induced implicitly within a machine learning framework that is trained jointly for

multiple language pairs. In Schwenk and Douze (2017), encoders and decoders for the given languages are trained jointly using a neural sequence to sequence model (Sutskever et al., 2014) using parallel corpora that are partially aligned; that is, each language within a pair is also part of at least one other parallel corpus. Neural machine translation can also be achieved with a single encoder and decoder that handles several input languages (Johnson et al., 2017), but the latter has not been evaluated as a general-purpose sentence representation model. According to Hill et al. (2016), the quality of the representations induced using a machine translation objective is lower than other neural models trained with different compositional objectives, such as Denoising Auto-Encoders and Skip-Thought (Kiros et al., 2015). Mono-lingual evaluation of sentence representation models can be found in Hill et al. (2016), Aldarmaki and Diab (2018), and Conneau and Kiela (2018). In Aldarmaki and Diab (2016), a modular training objective has been proposed for cross-lingual sentence embedding. However, their application was limited to the specific matrix factorization model they discussed. More recently, Conneau et al. (2018) proposed a modular transfer learning objective and evaluated it on neural sentence encoders using cross-lingual natural language inference classification. Our representation transfer framework is very similar to their approach, although we use a simpler loss function. In addition, we evaluate the framework as a general-purpose sentence encoder and compare it to other frameworks.

3 Approach

We selected two modern general-purpose sentence embedding models, the Inference-based classification model (`InferSent`) described in Conneau et al. (2017a), and the Sequential Denoising Auto-Encoder (`SDAE`) described in Hill et al. (2016). Both are implemented using a bidirectional LSTM network as an encoder followed by a classification or decoding network. We describe three possible cross-lingual alignment frameworks:

Joint cross-lingual modeling: We extend the monolingual objective of each model to multiple languages to be trained simultaneously via direct cross-lingual interactions in the objective function. This is in line with most existing cross-lingual extensions for top-down compositional models

Representation transfer learning: We directly optimize the sentence embeddings of new languages to match their translations in a parallel language (i.e. English). A similar approach was independently developed in Conneau et al. (2018).

Sentence mapping: Following the modular alignment framework for word embeddings (Smith et al., 2017), we fit an orthogonal transformation matrix on monolingual embeddings using a parallel corpus as a dictionary. Sentence mapping has been evaluated for word averaging models in Aldarmaki and Diab (2019).

3.1 Architectures

Most neural sentence embedding models are based on a sequential encoder—typically a bi-directional Long Short-Term Memory (Schuster and Paliwal, 1997)—followed by either a sequential decoder or a classifier. These models can be categorized according to their training objective:

Classification Accuracy: Sentence encoders can be trained by maximizing the accuracy in an extrinsic evaluation task. For example, InferSent (Conneau et al., 2017a) is trained on the Stanford Natural Language Inference (SNLI) dataset for inference classification (Bowman et al., 2015sss). This type of model requires labeled training data, which can make it challenging to expand across different languages.

Reconstruction: Using raw monolingual data, sentence encoders can be trained by minimizing the reconstruction loss, where a decoder is trained simultaneously to reconstruct the input sentence from the intermediate representation—e.g. Sequential Auto-Encoder (SAE) and Sequential Denoising Auto-Encoder (SDAE) (Hill et al., 2016). The latter introduces textual noise on the input sentence to make the embeddings more robust.

Translation: In Neural Machine Translation (NMT), a model is trained to maximizes the accuracy of generating a translation from the intermediate representation of the source sentence. Unlike modern NMT systems that rely on attention mechanisms, this model is trained for the purpose of sentence embedding, so only the intermediate representations are used as input to the decoder. This model requires parallel corpora for training.

The three objectives above are illustrated in Figures 1 and 2. We use the single-layer bidirectional LSTM encoder architecture with max-pooling described in Conneau et al. (2017a) for all encoders, and an LSTM decoders for SDAE and NMT.

(a) Unrolled LSTM encoder
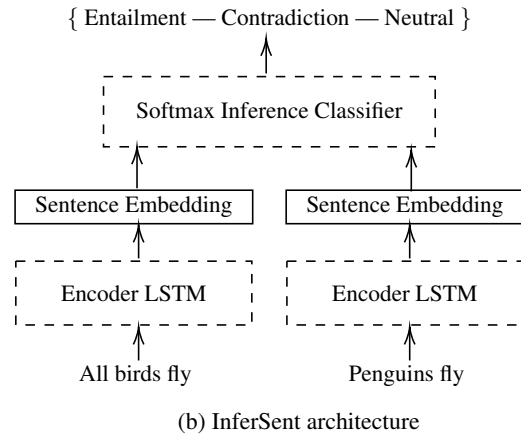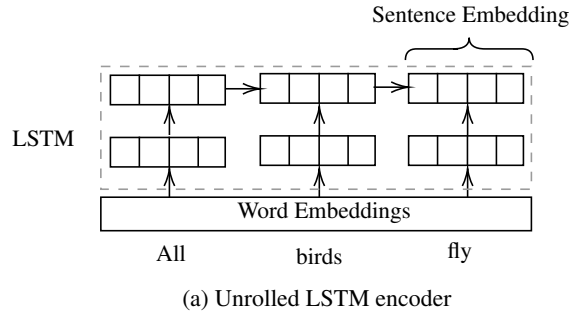

(b) InferSent architecture

Figure 1: Illustrations of neural sentence embedding architectures based on LSTM encoders. (a) shows an unrolled LSTM encoder with word embeddings. (b) shows InferSent architecture with a softmax classification network on top of the encoder.
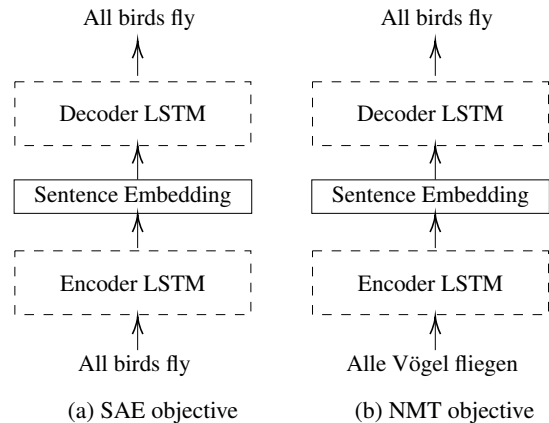

(a) SAE objective   (b) NMT objective

Figure 2: Illustrations of LSTM encoder-decoder architectures for sentence embeddings. (a) Sequential Auto-Encoder objective, where the input and output are the same sentence. (b) Neural Machine Translation objective, where the output is a translation of the input sentence from a parallel corpus.

3.2 Joint Cross-Lingual Modeling

We first discuss our joint cross-lingual neural models based on the above architectures. Note that joint modeling requires modifying the architecture and objective function of each model in a way that includes simultaneous interactions of cross-lingual sentence embeddings. This can be achieved in various ways with any degree of complexity, but we specifically aim to evaluate a direct extension of each loss function without extraneous objectives or constraints.

3.2.1 Joint Cross-Lingual Encoder-Decoder

The Sequential Denoising Auto-Encoder (`SDAE`) is trained to reconstruct the original input sentence from the intermediate sentence representation, where the input is corrupted with linguistic noise, such as word substitutions and reordering (Hill et al., 2016). This allows the model to robustly learn sentence representations from raw monolingual data. The Neural Machine Translation model, as depicted in Figure 2, has an identical architecture, with the only difference being the language of the input sentence. A cross-lingual extension of `SDAE` naturally leads to the `NMT` objective. We combine the `SDAE` and `NMT` objectives in a joint architecture, where multiple encoders are trained simultaneously with a single shared decoder. We alternate the input language (and the encoder) in each training batch, and the intermediate sentence embeddings are used as input to the shared decoder. Since the decoder is trained to predict the target sentence from the intermediate sentence representation regardless of input language identity, the encoders are expected to be updated in a way that results in consistent cross-lingual embeddings. Joint multi-lingual NMT has been previously shown to yield cross-lingual representations, as in Schwenk and Douze (2017).

3.3 Joint Cross-Lingual InferSent

Since `InferSent` is trained with an extrinsic classification objective, bilingual or multilingual optimization requires annotated data in each language. At the time of development, the SNLI dataset was only available in English[1], so we translated the training and evaluation datasets to Spanish and German using Amazon Translate. Note that in practice, machine translation might

---

[1]Other cross-lingual natural language inference corpora are now publicly available (Conneau et al., 2018), but our experiments were conducted before their release.

not be a viable option, especially if we try to extend the model to low-resource languages. Modern `NMT` systems require millions of parallel sentences to achieve good translation performance. For our purposes, the translated data allow us to assess the performance in different settings.
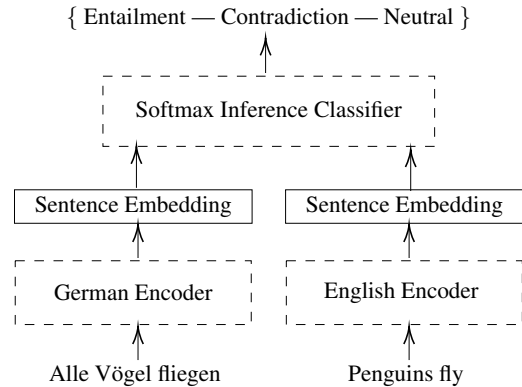
Figure 3: Illustrations of a joint training step, where different languages are used for the premise and hypothesis.

Similar to the joint `SDAE`/`NMT` model, we train encoders for all languages simultaneously. Since the input to the classifier consists of an ordered pair of sentences, we randomly pick a language for the premise and a language for the hypothesis in each training batch and use their respective encoders. A single classifier is shared regardless of the input languages. Similar to the monolingual case, the model is trained to maximize the performance in the inference classification task, which is cross-lingual in this case. An illustration of a training example is shown in Figure 3, where the premise is in German and the hypothesis in English.

3.4 Representation Transfer Learning

In the representation transfer framework, we use a monolingual pre-trained model to guide the training of additional encoders without the original supervised training objective. Using a parallel corpus that has source sentences aligned with English translations, we first generate the representations for the English sentences using a pre-trained `SDAE` or `InferSent` model. Then, we use these representations as a target to train an encoder for the other language in a supervised manner. The pivot encoder remains unchanged and only the new encoder is updated during training to ensure that independently trained encoders will still be aligned. Several functions can be used to achieve this, such as the L1 or L2 loss to minimize the distances be-
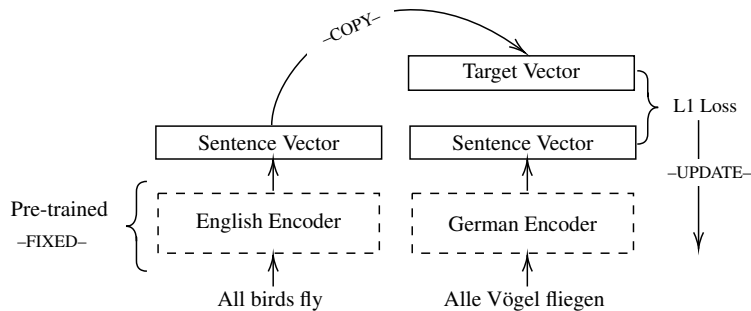
Figure 4: Representation transfer model, with pre-trained English encoder and L1 loss.

tween the source and target representations, or to maximize the cosine of the angle between them. Empirically, we observed no notable difference between these alternatives.[2] The transfer learning approach is illustrated in Figure 4.

3.5 Sentence Mapping

We follow the approach used for word-level transformation, where a dictionary is used to fit an orthogonal transformation matrix from the source to the target vector space (Smith et al., 2017). To extend this to sentences, we use a parallel corpus as a dictionary, and fit a transformation matrix between their sentence embeddings. After training, we apply the learned transformation post-hoc on newly generated sentence embeddings.

4 Evaluation

In a well-aligned cross-lingual vector space, sentences should be clustered with their translations across various languages. As discussed in Schwenk and Douze (2017) this can be measured with sentence translation retrieval: the accuracy of retrieving the correct translation for each source sentence from the target side of a test parallel corpus. This is done using nearest neighbor search with the cosine as a similarity measure. While not exactly an intrinsic evaluation metric, this scheme is the closest measure of alignment quality at the sentence level across all features in the vector space.

We used bottom-up embeddings composed using weighted averaging with smooth inverse frequency (Arora et al., 2017; Aldarmaki and Diab, 2018), which has been shown to work well as monolingual sentence embeddings compared to

other bottom-up approaches. We use skipgram with subword information (Bojanowski et al., 2017), i.e. `FastText`, for the word embeddings, which are also used as input to the neural models. We applied static dictionary alignment using the approach and dictionaries in Smith et al. (2017), in addition to sentence mapping using the parallel corpora. We trained the monolingual `FastText` word embeddings and SDAE models using the 1 Billion Word benchmark (Chelba et al., 2014) for English, and WMT'12 News Crawl data for Spanish and German (Callison-Burch et al., 2012). We used WMT'12 Common Crawl data for cross-lingual alignment, and WMT'12 test sets for evaluations. We used the augmented SNLI data described in (Dasgupta et al., 2018) and their translations for training the mono-lingual and joint `InferSent` models. For all datasets and languages, the only preprocessing performed was tokenization.

One of our evaluation objective is to assess the minimal bilingual data requirements for each framework, so we split the training parallel corpora into subsets of increasing size from 1,000 to 1 million sentences, where we double the size in each step. We report sentence translation retrieval accuracies in all language directions, using *en* for English, *es* for Spanish, and *de* for German [3].

4.1 Results

The results of the various SDAE models compared with the baselines are shown in Figure 5. With less than 100K parallel sentences, the joint `SDAE/NMT` model yielded poor performance compared to all models, but with 100K and more

---

[2]We settled on using Adam optimization (Kingma and Ba, 2014) with L1 loss.

[3]This evaluation scheme was recently introduced in Aldarmaki and Diab (2019) with data splits that are now available for download. Note that we used slightly older datasets in our experiments.
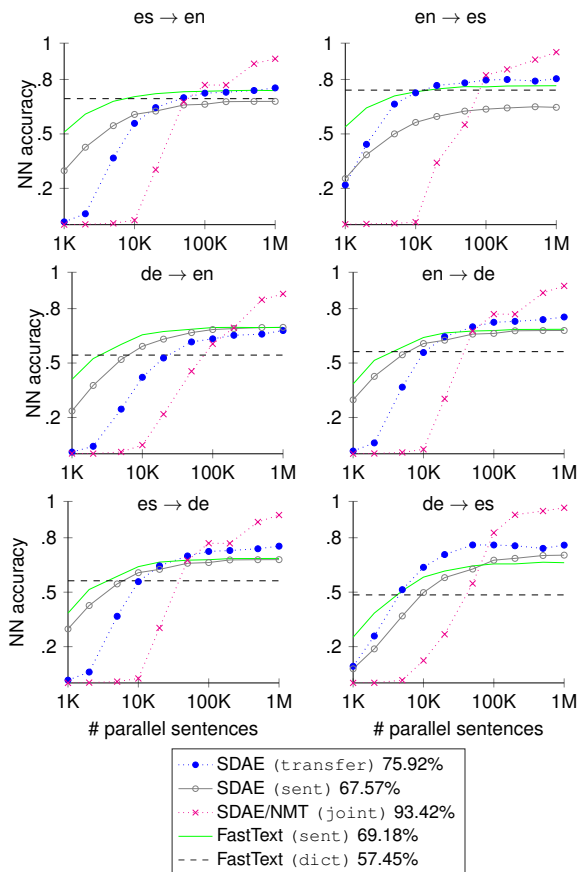
Figure 5: Nearest neighbor translation accuracy as a function of (log) parallel corpus size. (sent) to sentence-level mapping, and (dict) refers to the baseline (using a static dictionary for mapping). The legend shows the average accuracies of each model using 1M parallel sentences.

Figure 6: Nearest neighbor translation accuracy as a function of (log) parallel corpus size. (sent) to sentence-level mapping, and (dict) refers to the baseline (using a static dictionary for mapping). The legend shows the average accuracies of each model using 1M parallel sentences.

data, the model quickly exceeded the performance of all others by a large margin. Transfer learning achieved the second best performance, although it lagged behind the joint model with large parallel sets. With small amounts of parallel text, all models outperformed the joint SDAE/NMT, particularly the word based FastText models. Sentence mapping performed on average better than the static dictionary baseline, but FastText sentence mapping was generally better.

Figure 6 shows the results of the InferSent alignment models. Note that the joint InferSent model was trained with supervision using the translated SNLI data instead of the variable-size parallel corpora, so the performance is constant with respect to the number of parallel sentences. The joint model did not learn to align the cross-lingual sentences. Possible explanations of this failure are discussed in section 4.3.

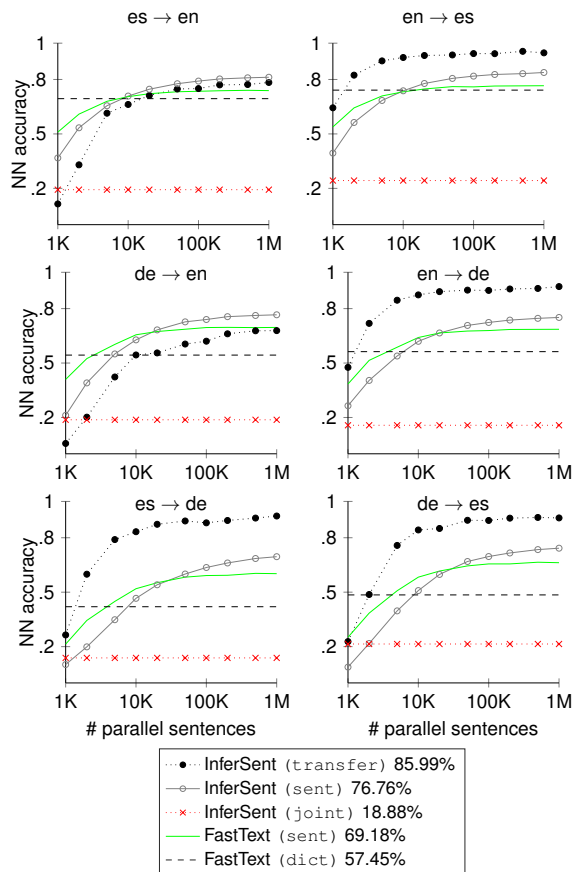Overall, the transfer learning model worked well for InferSent resulting in high transla-

tion retrieval accuracies even with relatively small amounts of parallel text ($\sim$ 5K sentences). Sentence mapping also performed better than the word-based baselines with additional parallel data (> 20K).

4.2 Overall Evaluation

In this section, we compare the overall performance of different types of models on sentence translation retrieval. We plotted the average cross-lingual accuracy (averaged over all language directions) by the best performing variant of each model in Figure 7. With small amounts of parallel text, around 5K sentences, the best performance was achieved using InferSent transfer model. The model continued to yield the highest performance until it was exceeded by the joint SDAE/NMT model at 500K sentences. The representation transfer models for SDAE exceeded the FastText model at around 20K sentences, and achieved comparable performance to InferSent sentence mapping.
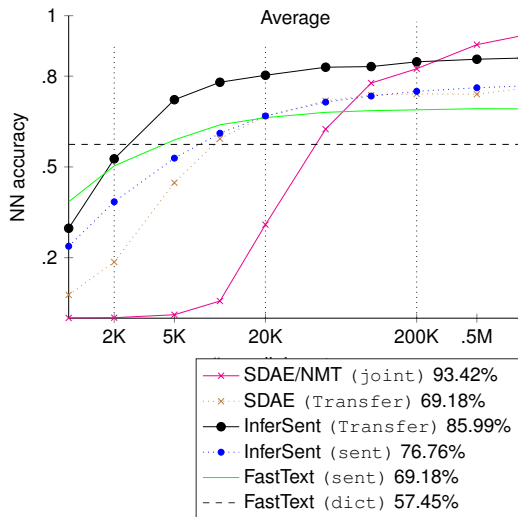
Figure 7: Nearest neighbor translation accuracy as a function of (log) parallel corpus size. The legend shows the average accuracies of each model using 1M parallel sentences.

Language	Monolingual Nearest Neighbors
Query: Tons of people are gathered around the statue	
Spanish	There are **several people sitting around** a table. There are **several people** outside of a building. **There are multiple people** present.
English	The **people** are taking photos of **the statue**. **A group of people** looking at **a statue**. **People are gathered** by the water.
Query: A vehicle is crossing a river	
Spanish	**A sedan** is stuck in the middle of **a river**. People are **crossing a river**. A **taxi cab** is **driving down a path of snow**.
English	A person is near **a river**. People are **crossing a river**. A **Land Rover** is splashing water as it **crosses a river**.

Table 1: Mono-lingual nearest neighbors (or their translations) of a sample of query sentences from SNLI test set using joint `InferSent` encoders. Phrases similar to the query sentences are shown in **bold**.

4.3 Analysis of Joint InferSent Performance

The joint `InferSent` model was trained to maximize the cross-lingual classification accuracy on cross-lingual inference data. The cross-lingual inference classification performance was comparable to the monolingual case for each language. The monolingual accuracies were around 83%, 79%, and 79% for English, German, and Spanish, respectively. The cross-lingual accuracy was around 79%. Given this relatively high performance in NLI classification and the poor performance in cross-lingual translation retrieval, we surmise that the 3-way classification objective is not demanding enough to learn general-purpose semantic representations. In addition, high per-

Language	Cross-lingual Nearest Neighbors
Query: Tons of people are gathered around the statue	
Spanish	Food and wine are on the table that has **many people surrounding** it. **Some people** enjoying their brunch together in the outdoor seating area of a restaurant... The **group of people** are game developers creating a new video game in their office.
English	The **group of people** are flying in the air on their unicorns . A **group of people** are standing around with smiles on their faces... A **group of people** dressed as clowns stroll into the Bigtop Circus holding signs.
Query: A vehicle is crossing a river	
Spanish	People and a baby are **crossing** the street at a crosswalk to get home. The person in the picture is riding **a bike slowing up hill** , pumping the pedals as hard as they can. The man , wearing scuba gear , jumps off the side of **the boat into the ocean** below.
English	A person in a coat with a briefcase **walks down a street** next to the bus lane. A man waterskiing in **a river** with a large wall in the background. A person **waterskiing in a river** with a wall in the background.

Table 2: Cross-lingual nearest neighbors (or their translations) of a sample of query sentences from SNLI test set using joint `InferSent` encoders. Phrases similar to the query sentences are shown in **bold**.

formance in a specific extrinsic evaluation task is not necessarily an indication of general embedding quality.

Tables 1 and 2 show examples of monolingual and cross-lingual nearest neighbors (or their English translations) from the hypotheses in SNLI test sets. The cross-lingual nearest neighbors did share several semantic aspects with the query sentence; subjects or verbs or combinations of these were observed in nearest neighbors. However, the exact translations were not the nearest neighbors in most cases, and the nearest neighbors often included several extraneous pieces of content not present in the query sentence. The mono-lingual nearest neighbors, on the other hand, were more semantically similar to each other, not only in the semantic features that are present, but also in their exclusions of dissimilar details.

We surmise that only a subset of semantic features were learned by the `InferSent` objective given the specific characteristics of the SNLI training sets. In other words, the model was not pushed to preserve the full semantic content since only a small subset of features were useful for entailment

relationships. The higher similarity among monolingual nearest neighbors is likely an artifact of the underlying word embeddings passing through the same encoder network.

4.4 Extrinsic Evaluation

Relying on a single measure is never sufficient to probe all characteristics of a vector space. Extrinsic evaluation can be another useful tool to measure the effectiveness of various cross-lingual models, although extrinsic tasks typically measure specific and narrow aspects of semantics. Nevertheless, we can still gain some insights about certain characteristics of these models and their applicability. One of the most widely used tasks for cross-lingual evaluation is the Cross-Lingual Document Classification benchmark (CLDC), where a model is trained in one language and tested on another (Schwenk and Li, 2018; Klementiev et al., 2012).

We report the average classification accuracies in CLDC across all language directions (a total of six directions) using the datasets in Schwenk and Li (2018); the multi-layer perceptron was used as a classifier trained for each source language, then tested in the remaining two.

The highest accuracy was achieved using `FastText` vectors, followed by `InferSent` transfer and sentence mapping models. With large enough parallel corpora, the performance of `SDAE/NMT` exceeded the transfer model, but with smaller data, `SDAE` transfer model achieved consistently higher performance.

These results are consistent with the trend of these models in mono-lingual topic categorization (Aldarmaki and Diab, 2018), where word averaging achieved consistently higher performance than all neural models. This indicates that cross-lingual models share the same semantic characteristics as their underlying mono-lingual counterparts. We should underscore that CLDC is a rather coarse categorization task where documents are classified into four categories. Note also that the `FastText` model achieved relatively high performance even when it was aligned with only 1K parallel sentences, a condition in which sentence translation retrieval accuracy was less that 40%. This poor correlation with sentence translation retrieval accuracies indicates that neither evaluation framework is reliable on its own. Our intuition is that sentence translation retrieval is a more com-
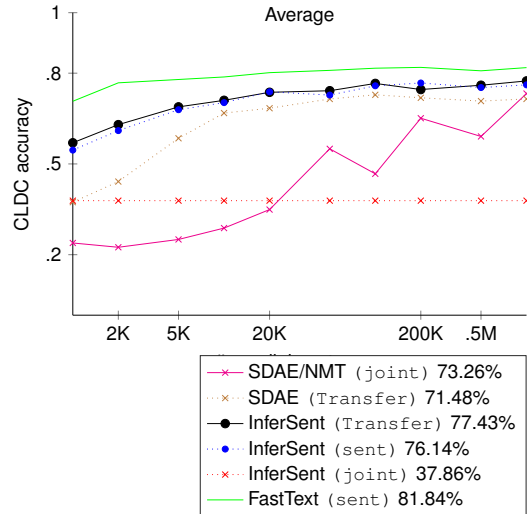

Figure 8: Average cross-lingual document classification accuracy as a function of (log) parallel corpus size. The legend shows the average accuracies of each model using 1M parallel sentences.

prehensive measure since all features in the vector space weigh equally in calculating the cosine similarity; on the other hand, a supervised classifier weighs features according to their correlations with the target classes.

5 Conclusions

We explored different approaches for cross-lingual alignment of top-down sentence embedding models: joint modeling, representation transfer, and sentence mapping. With sufficient amounts of parallel text, joint modeling yielded superior performance in the joint SDAE and NMT model, while joint InferSent failed to yield good alignments. Our results underscore the difficulty of joint modeling itself in addition to its relatively high data and memory requirements. With smaller amounts of parallel text, representation transfer worked reasonably well across all models, whereas sentence mapping was generally worse. Moreover, the transfer and sentence mapping frameworks enable modular training where additional languages can be added without retraining existing models and without labeled training data (as in InferSent), which allows scaling neural models to more languages with less resources. In extrinsic evaluation using cross-lingual document classification, transfer models achieved consistently better performance than joint models. Between the two sentence embedding models we evaluated, InferSent yielded better performance than SDAE and NMT, except in the joint framework.

In practice, joint and transfer learning can be combined in various ways according to data availability and modeling choices. A multi-task framework can be used to optimize both objectives at once. Given the lower data cost of representation transfer models, a joint model can be trained first for a set of resource-rich languages, followed by transfer learning for low-resource languages.

References

Hanan Aldarmaki and Mona Diab. 2016. Learning cross-lingual representations with matrix factorization. In *Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP*, pages 1–9.

Hanan Aldarmaki and Mona Diab. 2018. Evaluation of unsupervised compositional representations. *Proceedings of the 27th International Conference on Computational Linguistics*.

Hanan Aldarmaki and Mona Diab. 2019. Context-aware crosslingual mapping. *arXiv preprint arXiv:1903.03243*.

Hanan Aldarmaki, Mahesh Mohan, and Mona Diab. 2018. Unsupervised word mapping using structural similarities in monolingual embeddings. *Transactions of the Association of Computational Linguistics*, 6.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015sss. A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

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

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017a. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 748–756.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-hsuan Sung, Brian Strope, et al. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of NAACL-HLT*, pages 1367–1377.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association of Computational Linguistics*, 5(1):339–351.

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. *Proceedings of COLING 2012*, pages 1459–1474.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Hieu Pham, Thang Luong, and Christopher Manning. 2015. Learning distributed representations for multilingual text sequences. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 88–94.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

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)*, volume 2, pages 567–572.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

Hubert Soyer, Pontus Stenetorp, and Akiko Aizawa. 2014. Leveraging monolingual data for crosslingual compositional word representations. *arXiv preprint arXiv:1412.6334*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1403–1412.