LEARNING CROSS-LINGUAL SENTENCE REPRESENTATIONS VIA A MULTI-TASK DUAL-ENCODER MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural language models have been shown to achieve an impressive level of performance on a number of language processing tasks. The majority of these models, however, are limited to producing predictions for only English texts due to limited amounts of labeled data available in other languages. One potential method for overcoming this issue is learning cross-lingual text representations that can be used to transfer the performance from training on English tasks to non-English tasks, despite little to no task-specific non-English data. In this paper, we explore a natural setup for learning cross-lingual sentence representations: the dual-encoder. We provide a comprehensive evaluation of our cross-lingual representations on a number of monolingual, cross-lingual, and zero-shot/few-shot learning tasks, and also give an analysis of different learned cross-lingual embedding spaces.

1 INTRODUCTION

There has been a significant amount of recent work on developing models that can produce sentence representations that are useful for a number of language processing tasks (Kiros et al., 2015; Conneau et al., 2017; Subramanian et al., 2018; Logeswaran & Lee, 2018; Cer et al., 2018). However, these models are trained on largely monolingual data, and can thus only be used for tasks in a single language. A promising direction for extending the previous models to multiple languages is learning cross-lingual embedding spaces (Schwenk et al., 2017; Eriguchi et al., 2018; Singla et al., 2018), which could be used to transfer performance in one language to others.

We develop a novel approach for cross-lingual representation learning by combining the dual-encoder architectures used for learning sentence representations (Logeswaran & Lee, 2018; Cer et al., 2018) and for bi-text retrieval (Guo et al., 2018). By doing so, we learn representations that maintain state-of-the-art performance in tasks for a source language while *simultaneously* obtaining state-of-the-art performance in zero-shot learning tasks for a target language. For a given source-target language pair, we construct a multi-task training scheme using native source language tasks, native target language tasks, and a bridging source-target translation task to learn sentence representations that are aligned between the source and target languages. We then evaluate the learned representations on several monolingual and cross-lingual tasks, and also provide a graph-based analysis of the learned representations.

We find that multi-task training using additional monolingual tasks improves performance over models that only make use of parallel data on both cross-lingual semantic textual similarity (STS) (Cer et al., 2017) and Søgaard et al. (2018)'s cross-lingual eigen-similarity metric. The results show that the addition of monolingual data actually improves the embedding alignment of sentences and their translations. Furthermore, we find that cross-lingual training with additional monolingual data leads to far better transfer learning performance, and we show that our cross-lingual representations outperform state-of-the-art zero-shot learning systems in sentiment classification and natural language inference.

2 MULTI-TASK DUAL-ENCODER MODEL

The core of our approach is the idea of modeling various tasks as ranking input-response pairs by encoding them via two encoders, with the crucial task for learning cross-lingual representations

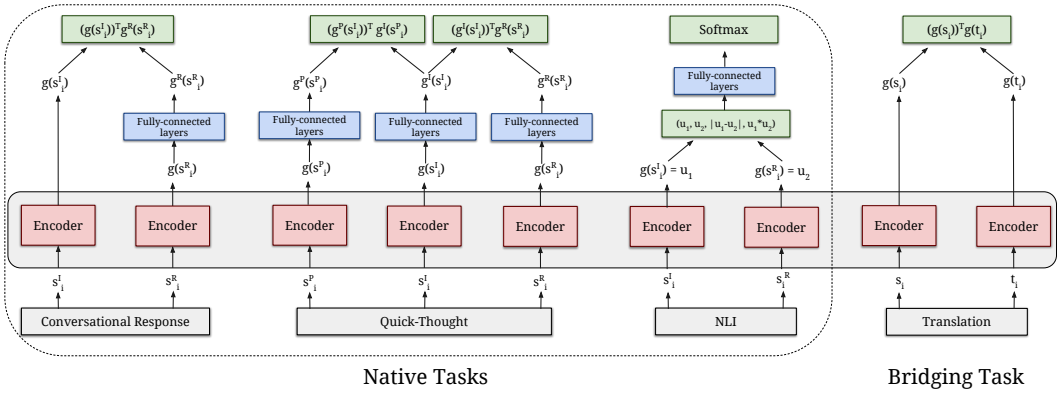

Figure 1: Multi-task dual-encoder model. It consists of a group of native tasks in each language and a bridging task using translation pair data. The encoders in the gray box all share their parameters, and thus constitute g. The Quick Thought task can be treated as a variation of the dual-encoder model, where we combine sentence-predecessor and sentence-successor models.

being *translation ranking*. For translation ranking, as well as for our other tasks, we take an input sentence s_i^I and an associated response sentence s_i^R, and we seek to rank s_i^R over all other possible response sentences $s_j^R \in \mathcal{S}^R$. To do so, we model the conditional probability $P(s_i^R \mid s_i^I)$ as:

$$P(s_i^R \mid s_i^I) = \frac{e^{\phi(s_i^I, s_i^R)}}{\sum_{s_j^R \in \mathcal{S}^R} e^{\phi(s_i^R, s_j^R)}}, \quad \phi(s_i^I, s_j^R) = g^I(s_i^I)^\top g^R(s_j^R) \tag{1}$$

Where g^I and g^R are the input and response sentence encoding functions that compose the dual-encoder. Since the normalization term in equation 1 is computationally intractable, we follow the approaches of Henderson et al. (2017) and instead choose to model an approximate conditional probability $\widetilde{P}(s_i^R \mid s_i^I)$:

$$\widetilde{P}(s_i^R \mid s_i^I) = \frac{e^{\phi(s_i^I, s_i^R)}}{\sum_{j=1, j \neq i}^{K} e^{\phi(s_i^R, s_j^R)}} \tag{2}$$

Where K denotes the size of a single batch of training examples, and the s_j^R correspond to the response sentences associated with the other input sentences in the same batch as s_i^I. We parametrize g^I and g^R as deep neural networks that are trained to minimize the negative log-likelihood of $\widetilde{P}(s_i^R \mid s_i^I)$ for each task.

In order to produce a single sentence encoding function g that can be evaluated on downstream tasks, we share several layers between the input and response encoders and treat the final output of these shared layers as g. Additionally, these layers are modeled after the Universal Sentence Encoder (USE) model of Cer et al. (2018), since it is the state-of-the-art model that is most amenable to our setup. To learn cross-lingual representations, we train g on several tasks from mirrored corpora across languages[1] for the source-target language pairs English-French (en-fr), English-Spanish (en-es), and English-German (en-de). The resulting model structure is illustrated in Figure 1.

2.1 Encoder Architecture

Word and Character Embeddings. As part of the training process for learning the cross-lingual sentence encoding function g, we learn embeddings for the words and characters present in the training data for a given source-target language pair. Word embeddings are learned end-to-end, as we noticed that pre-trained embeddings did not make a difference for final performance. Character embeddings are learned in a similar manner, but with the added stipulation that we consider character

[1] We explore some training setups without mirroring in the supplementary material.

n-gram embeddings instead of single character embeddings by using a single feedforward layer with $tanh$ activation on top of character n-grams. Each word in an input sentence then obtains a character embedding representation by having its character n-gram representations summed together. To have the sentence encoder g leverage the word and character embeddings together without drastically increasing its number of parameters, we sum the word and character embeddings before using them as input to g.

Transformer Encoder. The actual architecture of the shared encoder g consists of three[2] layers of transformer stacks, which contain the feed-forward and multi-head attention sub-layers described in Vaswani et al. (2017). The transformer encoder output is a variable-length sequence at each stack. We average encodings of all sequence positions in the final layer as the final sentence encoding. This embedding is then fed into different sets of feedforward layers that are used for each task. For our transformer layers, we use 8 attentions heads, a hidden size of 512, and a filter size of 2048 in all experiments.

2.2 Multi-task Training Setup

To learn a function g that is capable of strong cross-lingual matching and transfer learning performance for a source-target language pair while also maintaining monolingual downstream task performance, we employ four unique task types for each language pair. Specifically, we employ *a conversation response prediction task, a quick thought task, a natural language inference task*, and a bridging task – *translation ranking*. Six total tasks are used in training, as the first two tasks are mirrored across languages.

Conversation Response Prediction. We model the conversation response prediction task in the same manner as Yang et al. (2018). We minimize the negative log-likelihood of $\widetilde{P}(s_i^R \mid s_i^I)$, where s_i^I is a single comment and s_i^R is its associated response comment. For the response side, we model $g^R(s_i^R)$ as two fully-connected feedforward layers of size 320 and 512 with $tanh$ activation on top of $g(s_i^R)$. For the input side, however, we simply let $g^I(s_i^I) = g(s_i^I)$, as we noticed in early experiments that letting the optimization of the conversational response task more directly influence the parameters of the underlying sentence encoder g led to better downstream task performance.

Quick Thought. We use a modified version of the Quick Thought task detailed by Logeswaran & Lee (2018). We minimize the sum of the negative log-likelihoods of $\widetilde{P}(s_i^R \mid s_i^I)$ and $\widetilde{P}(s_i^P \mid s_i^I)$, where s_i^I is a sentence taken from an article and s_i^P and s_i^R are its predecessor and successor sentences respectively. For this task, we model all three of $g^P(s_i^P)$, $g^I(s_i^I)$, and $g^R(s_i^R)$ using separate, fully-connected feedforward layers of size 320 and 512 with $tanh$ activation on top of g, as we did for $g^R(s_i^R)$ in our conversational modeling task.

Natural Language Inference (NLI). We also include an *English-only* natural language inference task based on Bowman et al. (2015). For this task, we first encode an input sentence s_i^I and its corresponding response hypothesis s_i^R into vectors u_1 and u_2 using g. The vectors u_1, u_2 are then used to construct a feature vector $(u_1, u_2, |u_1 - u_2|, u_1 * u_2)$, where $(\cdot)$ represents concatenation and $*$ represents element-wise multiplication. The form of this feature vector is derived from the original experiments of Bowman et al. (2015). This feature vector is then fed into a single feedforward layer of size 512 that is used to perform the 3-way NLI classification.

Translation Ranking. Our translation task setup is identical to the one used by Guo et al. (2018) for bi-text retrieval. We minimize the negative log-likelihood of $\widetilde{P}(s_i \mid t_i)$, where (s_i, t_i) is a source-target translation pair. Since the translation task is intended to align the sentence representations for the source and target languages, we do not use any kind of task-specific feedforward layers and instead use g as both g^I and g^R. Following Guo et al. (2018), we append 5 translations that are similar to the correct translation to each training example as "hard-negatives". Similarity is determined via a version of our model trained only on the translation ranking task. We did not see additional gains from using more than 5 hard-negatives.

[2]We tried up to six layers, but did not notice a significant difference beyond three.

3 EXPERIMENTS

3.1 CORPORA

We draw upon multiple, openly available data sources and training corpora for the training of the tasks mentioned above. For each of our datasets, we use 90% of the data for training, and the remaining 10% for development/validation. Our data preprocessing procedures are described in detail in the supplementary material.

Reddit. We preprocess the Reddit data extracted by Al-Rfou et al. (2016) into 600 million input-response comment pairs for training our conversation response prediction task. We also translate this data using the Google neural machine translation (NMT) system of Wu et al. (2016).

Wikipedia. To get native, non-English data, we extract triplets of contiguous sentences from English, French, Spanish, and German articles take from Wikipedia. Our final extracted corpus of Wikipedia sentence triplets consists of 127.9, 49.5, 29.8, and 49.3 million triplets for English, French, Spanish, and German respectively, which we use to train our Quick Thought task.

Stanford Natural Language Inference (SNLI). The NLI data we use is taken from the Stanford Natural Language Inference (SNLI) dataset of Bowman et al. (2015), which consists of 570K sentence pairs associated with one of three labels: *entailment*, *contradiction*, or *neutral*. The corpus is split into training (550K), validation (10K), and testing sets (10K).

Translation. The data for training the translation task is constructed using a system similar to the approach described by Guo et al. (2018). The final constructed corpus contains around 600M en-fr pairs, 470M en-es pairs and 500M en-de pairs.

3.2 MODEL CONFIGURATION

In all of our experiments, multi-task training is done by cycling through the different tasks (translation pairs, Reddit, Wikipedia, NLI) and performing an optimization step for a single task at a time. We train all of our models with a batch size of 100 using stochastic gradient descent with a learning rate of 0.008. All of our models are trained for 30 million steps. All input text is tokenized prior to being used for training. We build a vocab containing 200 thousand unigram tokens with 10 thousand hash buckets for out-of-vocabulary tokens. The character n-gram vocab contains 200 thousand hash buckets used for 3 and 4 grams. Both the word and character n-gram embedding sizes are 320. All hyperparameters are tuned based on performance on the development portion (random 10% slice) of our datasets. Finally, as an additional training heuristic, we multiply the gradients to the word and character embeddings by a factor of 100[3]. We found that using this embedding gradient multiplier alleviated vanishing gradient issues and greatly improved training.

We compare the proposed cross-lingual multi-task (referred to simply as multi-task) models with baseline models that are trained using only the translation ranking task, which we dub as the "translation-ranking" models.

3.3 MODEL PERFORMANCE ON ENGLISH DOWNSTREAM TASKS

We first evaluated all of our cross-lingual models on several downstream English tasks taken from SentEval (Conneau & Kiela, 2018) to verify the impact of cross-lingual training. Each task is described in the supplementary material. Results on the tasks are summarized in Table 1. We note that cross-lingual training does not hinder the effectiveness of our encoder on English tasks, as the multi-task models are close to state-of-the-art in each of the downstream tasks. For the Text REtrieval Conference (TREC) eval, we actually find that our multi-task models outperform the previous state-of-the-art models by a sizable amount.

Table 1 also includes the results for our translation-ranking models on the same downstream English tasks. We include these results mainly to gauge the level of semantic information that can be learned from using only translation pair data. As expected, the performance of the translation-ranking models is significantly worse than that of the multi-task models.

[3]We tried different orders of magnitude for the multiplier and found 100 to work the best.

Table 1: Performance on classification transfer tasks.

Model	MR	CR	SUBJ	MPQA	TREC	SST	STS Bench (dev / test)
Cross-lingual Multi-task Models							
en-fr	77.9	82.9	95.5	89.3	95.3	84.0	0.803 / 0.763
en-es	80.1	85.9	94.6	86.5	96.2	85.2	0.809 / 0.770
en-de	78.8	84.0	95.9	87.6	96.1	85.0	0.802 / 0.764
Translation-ranking Models							
en-fr	68.7	79.3	87.0	81.8	89.4	74.2	0.668 / 0.558
en-es	67.7	75.7	83.5	86.0	94.4	72.6	0.669 / 0.631
en-de	67.8	75.2	84.4	83.6	86.8	74.6	0.673 / 0.632
State-of-the-art Models							
InferSent	81.1	86.3	92.4	90.2	88.2	84.6	0.801 / 0.758
Skip-Thought LN	79.4	83.1	93.7	89.3	–	–	–
Quick-Thought	82.4	86.0	94.8	90.2	92.4	87.6	–
USE Transformer	81.4	87.4	93.9	87.0	92.5	85.4	0.814 / 0.782

3.4 Cross-lingual Retrieval

We also evaluate both the multi-task and translation-ranking models' efficacy in performing cross-lingual retrieval by using held-out translation pair data. Following Henderson et al. (2017), we use precision at N (P@N) as the evaluation metric by checking if a source sentence's target translation ranks (where ranking is done using dot product) in the top N scored candidates when considering K other randomly selected target sentences. Unlike Henderson et al. (2017), we set K to be 999 instead of 99. This is because using $K = 99$ results in all metrics quickly shooting up to 99%, which leads us to believe that choosing the correct translation out of only 100 samples is too easy.

Table 2 summarizes the precision at 1 (P@1) metric of the multi-task models and translation-ranking models. The translation-ranking model remains as a strong baseline for finding the true translation, with up to 97.8% for the en-de retrieval task. The multi-task model performs almost identical to the translation-ranking model in all metrics, which provides some empirical justification that it is possible to maintain embedding space alignment despite optimizing for native tasks in each individual language. We also experimented with P@3 and P@10, the results are identical.

Table 2: Precision at 1 (P@1) result on a hold-out testing dataset for en-fr, en-es and en-de. Models attempt to predict the true translation target for a source sentence against 999 randomly targets.

	en-fr	en-es	en-de
Multi-task	95.4	87.5	97.5
Translation-ranking	95.1	88.8	97.8

3.5 Multilingual STS

We further test whether our learned cross-lingual representations can also perform well in their associated non-English language tasks by evaluating semantic textual similarity (STS) performance on French, Spanish, and German.

To evaluate Spanish-Spanish (es-es) STS, we use SemEval-2017 task 1 (STS17) track 3 of Cer et al. (2017), which contains 250 Spanish sentence pairs with human labeled similarity scores. We also evaluate Spanish-English (es-en) STS by using the track 4(a) task[4], which contains 250 en-es sentence pairs.

Beyond English and Spanish, however, there are no standard STS datasets available for other languages. As such, we evaluate on a translated version of the STS Benchmark dataset from Cer et al.

[4]The es-en task is split into track 4(a) and track 4(b), we only use track 4(a) here. The track 4(b) task contains sentence pairs from WMT with only one annotator for each pair. The previous reported numbers are also particularly low for this task. We suspect it is not a very good evaluation set for current systems.

(2017) for French, Spanish, and German. We use Google's translation system to translate the STS Benchmark sentences to French, Spanish and German. We believe that the results on our pseudo-multilingual STS Benchmark dataset are expected to still be a reasonable indicator of multilingual semantic similarly performance, since the NMT encoder-decoder architecture for translation differs significantly from our dual-encoder approach.

Following Cer et al. (2018), we first compute the sentence encodings u, v of an STS sentence pair, and then score the sentence pair similarity based on the angular distance between the two vectors, $-\arccos\left(\frac{uv}{||u||\,||v||}\right)$. Table 3 shows the Pearson's correlation coefficient of the STS tasks for all models. The first column shows the trained model performance on original English STS Benchmark data. Columns 2 to 4 shows the the performance on the other languages. All multi-task models remain strong on the translated STS tasks, with around 0.77 for dev and 0.74 for test in all languages. Lastly, columns 5 and 6 shows the results of en-es models on STS17 tasks. The un-tuned multi-task models achieve 0.827 for the es-es task and 0.769 for the es-en task. As a point of reference, we also list the two best performing STS systems, Tian et al. (2017) (ECNU) and Wu et al. (2017) (BIT), reported from Cer et al. (2017). Our results are very close to these state-of-the-art feature engineered and mixed systems, which we describe in greater detail in the supplementary material.

Table 3: Pearson's correlation coefficients on translated STS Benchmark and STS17 tasks. The first column shows the results on the original STS Benchmark data in English.

Model	**Translated STS Benchmark (dev / test)**				**STS17**	
	en-en	**fr-fr**	**es-es**	**de-de**	**es-es**	**es-en**
Multi-task en-fr	0.803 / 0.763	0.777 / 0.738	–	–	–	–
Trans.-ranking en-fr	0.668 / 0.558	0.641 / 0.579	–	–	–	–
Multi-task en-es	0.809 / 0.770	–	0.779 / 0.744	–	0.827	0.769
Trans.-ranking en-es	0.669 / 0.631	–	0.622 / 0.611	–	0.642	0.587
Multi-task en-de	0.802 / 0.764	–	–	0.768 / 0.722	–	–
Trans.-ranking en-de	0.673 / 0.632	–	–	0.630 / 0.526	–	–
ECNU	–	–	–	–	0.856	0.813
BIT	–	–	–	–	0.846	0.749

4 ZERO-SHOT CLASSIFICATION

To evaluate the transfer learning capabilities of our models, we examine how well the multi-task and translation-ranking encoders perform on zero-shot and few-shot classification tasks.

4.1 MULTILINGUAL NLI

We first evaluate the zero-shot classification performance of our multi-task models on multilingual natural language inference (NLI) tasks. We make use of the professionally translated French and Spanish subsets of SNLI created by Agić & Schluter (2017) for the cross-lingual zero-shot evaluation. There are 1000 examples in the translated subsets for each language. To evaluate, we simply feed the French or Spanish examples into the pre-trained English NLI sub-network of our cross-lingual models.

Table 4 lists the accuracy on the English SNLI test set (10k sentence pairs) for all models and the accuracy on the French and Spanish translated subsets (1k sentence pairs) for our en-fr and en-es models. The original English SNLI accuracies are around 84% for all of our multi-task models, indicating that English SNLI performance remains stable in the multi-task training setting. The zero-shot accuracy on the translated subsets of SNLI are around 74% for both of French and Spanish.

Row 4 shows the zero-shot French NLI performance of Eriguchi et al. (2018), which is a state-of-the-art zero-shot NLI classifiers based on multilingual NMT embeddings. Our multi-task model shows comparable performance to the NMT-based model in both English and French.

4.2 AMAZON REVIEW

Zero-shot Learning. We also conduct a zero-shot evaluation based on the Amazon review data extracted by Prettenhofer & Stein (2010). We preprocess the Amazon reviews and convert the data

Table 4: Zero-shot classification accuracy (%) on SNLI dataset.

Model	en (10k)	fr (1k)	es (1k)
Multi-task en-fr	83.7	74.0	–
Multi-task en-es	83.6	–	74.5
Multi-task en-de	84.3	–	–
Eriguchi et al. (2018) (NMT en-fr)	84.4	73.9	–

into a sentiment classification task by considering reviews with strictly more than three stars as positive and strictly less than three stars as negative, in the same manner as Prettenhofer & Stein (2010). Each review contains a summary field and a text field, which we concatenate to produce a single input. As the multi-task models are trained with sentence lengths clipped to 64, we only take the first 64 tokens from the the concatenated text as the input. There are 6000 training reviews in English, which we split into 90% for training and 10% for development.

We first encode inputs using the pre-trained multi-task and translation-ranking encoders and feed the encoded vectors into a 2-layer feed-forward network culminating in a softmax layer. We use layers of size 512 and $tanh$ activation functions in each layer. We use Adam for optimization with an initial learning rate of 0.0005 and a learning rate decay of 0.9 at every epoch during training. We use a batch size of 16 and train for 20 total epochs in all experiments. We freeze the cross-lingual encoder during training. The model architecture and parameters are tuned on the development set.

We first train the classifier on English data, and then evaluate it on the 6000 French and German Amazon review test examples. The results are summarized in Table 5. The accuracy on the English test set is 87.4% for the en-fr model and 87.1% for the en-de model, with the zero-shot accuracy being above 80% for both models. The translation-ranking models again perform worse on all metrics. Once again we compare the proposed model with Eriguchi et al. (2018), and find that our zero-shot performance has a reasonable gain on the fr test set[5].

Table 5: Zero-shot sentiment classification accuracy(%) on target language Amazon review test data after training on only English Amazon review data.

Model	en	fr	de
Multi-task en-fr	87.4	82.3	–
Translation-ranking en-fr	74.4	66.3	–
Multi-task en-de	87.1	–	81.0
Translation-ranking en-de	73.8	–	67.0
Eriguchi et al. (2018) (NMT en-fr)	83.2	81.3	–

Few-shot Learning. We further evaluate the proposed multi-task models via few-shot learning, by training on English reviews and only a portion of French and German reviews. Our few-shot models are compared with baselines of training on French and German reviews only. Table 6 shows the classification accuracy of the few-shot models, where the second row shows the percent of French and German data that is used when training each model. With as little as 20% of the French or German training data, the few-shot models perform nearly as good as the baseline models trained on 100% of the French and German data. Adding more French and German training data leads to further improvements in few-shot model performance, with the few-shot models reaching 85.8% accuracy in French and 84.5% accuracy in German when using all of the French and German data.

5 Analysis of Cross-lingual Embedding Spaces

Motivated by the recent work of Søgaard et al. (2018) studying the graph structure of multilingual word representations, we perform a similar analysis for our learned cross-lingual sentence representations. To do so, we take N samples of size K from en-fr, en-es, and en-de translation data and then encode these samples using the corresponding multi-task and translation-ranking models. We then

[5]Eriguchi et al. (2018) also train a shallow classifier, but use only review text and truncate their inputs to 200 tokens. Our setup is slightly different, as our models can take a maximum of only 64 tokens.

Table 6: Sentiment classification accuracy(%) on target language Amazon review test data after training on English Amazon review data and a portion of French of German data. The second row shows the percent of French (fr) or German (de) data is used for training in each model.

Model	**fr**						**de**					
	0%	10%	20%	40%	80%	100%	0%	10%	20%	40%	80%	100%
Few-shot	82.3	84.4	84.4	84.8	85.2	85.8	81.0	81.6	83.3	84.0	84.7	84.5
Baseline	–	79.2	80.0	82.7	84.3	84.9	–	75.5	77.7	81.6	83.5	84.4

compute pairwise distance matrices within each sampled set of encodings, and use these distance matrices to construct graph Laplacians[6]. Finally, we obtain the similarity $\Psi(S,T)$ between each model's source and target language embedding subsets by comparing the eigenvalues of the source language graph Laplacians to the eigenvalues of the target language graph Laplacians as follows:

$$\Psi(S,T) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}(\lambda_j(L_i^{(s)}) - \lambda_j(L_i^{(t)}))^2 \tag{3}$$

Where $L_i^{(s)}$ and $L_i^{(t)}$ refer to the graph Laplacians of the source language and target language sentences obtained from the ith sample of source-target translation pairs. A smaller value of $\Psi(S,T)$ indicates higher eigen-similarity of the source language and target language embedding subsets. Following Søgaard et al. (2018) we use a sample size of $K = 10$ translation pairs, but we choose to use $N = 1000$ samples instead of $N = 10$ (as was done in their work) since we found $\Psi(S,T)$ to have very high variance at $N = 10$. The computed values of $\Psi(S,T)$ for our multi-task and translation-ranking models are summarized in Table 7.

Table 7: Average eigen-similarity values of source and target embedding subsets.

Model	**en-fr**	**en-es**	**en-de**
multi-task	0.592	0.526	0.761
translation-ranking	1.036	0.572	2.187

We find that the source and target embedding subsets constructed from the *multi-task models* exhibit greater average eigen-similarity than those resulting from the translation-ranking models for all source-target language pairs. This result is not necessarily intuitive, since one might expect the translation-ranking model to optimize more for alignment. Given that eigen-similarity correlates with the better performance of the multi-task models in almost all tasks, a potential direction for future work could be to introduce regularization penalties based on graph similarity in multi-task training. Interestingly, we also observe that the eigen-similarity gaps between the multi-task and translation-ranking models are not uniform across language pairs (although it may be that translation-ranking requires even more training). Thus, another direction could be to further study differences in the difficulty of aligning different source-target language embeddings.

6 Conclusion

In this work, we explored a straightforward framework for training cross-lingual, multi-task dual-encoder models. We showed that by training English-French, English-Spanish, and English-German multi-task models using our setup, we can achieve near-state-of-the-art or state-of-the-art performance in a variety of English tasks while also being able to produce similar caliber results in zero-shot transfer learning tasks for other languages. Finally, we note that the fact that multi-task training can actually improve performance on some downstream English tasks (TREC) is particularly interesting, and believe that there are many possibilities for future explorations of cross-lingual model training.

[6]See Zhang (2011) for an overview of graph Laplacians.

REFERENCES

Željko Agić and Natalie Schluter. Baselines and test data for cross-lingual inference. *arXiv preprint arXiv:1704.05347*, 2017.

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Conversational contextual cues: The case of personalization and history for response ranking. *CoRR*, abs/1606.00372, 2016. URL http://arxiv.org/abs/1606.00372.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642. Association for Computational Linguistics, 2015. doi: 10.18653/v1/D15-1075. URL http://www.aclweb.org/anthology/D15-1075.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S17-2001.

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. Universal sentence encoder. *CoRR*, abs/1803.11175, 2018. URL http://arxiv.org/abs/1803.11175.

Alexis Conneau and Douwe Kiela. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*, 2018.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 670–680. Association for Computational Linguistics, 2017. URL http://aclweb.org/anthology/D17-1070.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*, 2018.

Dan Gillick. Sentence boundary detection and the problem with the u.s. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pp. 241–244, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1620853.1620920.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernández Ábrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Effective parallel corpus mining using bilingual sentence embeddings. *CoRR*, abs/1807.11906, 2018.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652, 2017. URL http://arxiv.org/abs/1705.00652.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 168–177, New York, NY, USA, 2004. ACM. ISBN 1-58113-888-1. doi: 10.1145/1014052.1014073. URL http://doi.acm.org/10.1145/1014052.1014073.

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.

Xin Li and Dan Roth. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pp. 1–7. Association for Computational Linguistics, 2002.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJvJXZb0W.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity. In *Proceedings of ACL*, pp. 271–278, 2004.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pp. 115–124, 2005.

Peter Prettenhofer and Benno Stein. Cross-Language Text Classification using Structural Correspondence Learning. In *48th Annual Meeting of the Association of Computational Linguistics (ACL 10)*, pp. 1118–1127. Association for Computational Linguistics, July 2010. URL http://www.aclweb.org/anthology/P10-1114.

Holger Schwenk, Ke Tran, Orhan Firat, and Matthijs Douze. Learning joint multilingual sentence representations with neural machine translation. *CoRR*, abs/1704.04154, 2017. URL http://arxiv.org/abs/1704.04154.

Karan Singla, Dogan Can, and Shrikanth Narayanan. A multi-task approach to learning multilingual representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 214–220. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/P18-2035.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Anders Søgaard, Sebastian Ruder, and Ivan Vulic. On the limitations of unsupervised bilingual dictionary induction. *CoRR*, abs/1805.03620, 2018. URL http://arxiv.org/abs/1805.03620.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B18WgG-CZ.

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. Ecnu at semeval-2017 task 1: Leverage kernel-based traditional nlp features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 191–197, 2017.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pp. 1101–1109, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1873781.1873905.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

Hao Wu, Heyan Huang, Ping Jian, Yuhang Guo, and Chao Su. Bit at semeval-2017 task 1: Using semantic information space to evaluate semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 77–84, 2017.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL http://arxiv.org/abs/1609.08144.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-Yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pp. 164–174. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/W18-3022.

X.-D. Zhang. The Laplacian eigenvalues of graphs: a survey. *ArXiv e-prints*, November 2011.

A SUPPLEMENTARY MATERIAL

A.1 DATA PREPROCESSING

In order to effectively use Reddit, Wikipedia, and translation data, we do a significant amount of preprocessing on the raw data. We describe our preprocessing procedures for each dataset in the following paragraphs.

Reddit. The raw Reddit corpus extracted by Al-Rfou et al. (2016) consists of 2.4 billion posts and comments from Reddit between 2007 and 2016, making it our largest data source by an order of magnitude. To preprocess this data for training our conversation response prediction task, we follow the same procedure as Yang et al. (2018). Essentially, we consider Reddit comments and their children (response comments) as input-response pairs. We filter out comments that have ≥ 350 characters, due to the limitations on the number of input tokens our models can accept. We also remove comments that start with "https", "@", or "/r/", and also remove comments whose authors have "bot" in their usernames. Lastly, we find that Reddit comments also contain a small mix of non-English text, and we filter out comments where the percentage of alphabetic characters is $\leq 70\%$. As mentioned in the paper, the final, processed Reddit data consists of 600 million input-response pairs.

Since Reddit comments are predominantly English, we found the original data to be unsuitable for training encoders for non-English languages. To create conversational corpora for other languages, we translate the entire Reddit dataset using Google's neural machine translation (NMT) system of Wu et al. (2016).

Wikipedia. One concern with using the translated Reddit corpus as a monolingual task for non-English languages is that the translated data will propagate any errors made by the NMT system used for translation. To ameliorate this problem, we crawl Wikipedia and extract triplets of contiguous sentences from Wikipedia articles. We use Wikipedia due to it having well-formed articles with a broad coverage of several languages. We extract all Wikipedia articles for English, French, Spanish and German from the Wikipedia dump of May 5, 2018, and use the sentence splitter model of Gillick (2009) to split the articles into sentences. All article titles and article section names are treated as sentences inline. As mentioned in the paper, our final extracted Wikipedia corpus consists of 127.9 million English sentence triplets, 49.5 million French triplets, 29.8 million Spanish triplets, and 49.3 million German triplets for French, Spanish and German respectively.

Stanford Natural Language Inference (SNLI). We use the English Stanford Natural Language Inference (SNLI) dataset consisting of 570K sentence pairs of Bowman et al. (2015) as is, without any further preprocessing. We also use the provided splits for training (550K), validation (10K), and testing (10K).

Translation. The data for training the translation task is crawled from the public web using a system similar to the approach described by Uszkoreit et al. (2010). The extracted data is further cleaned by a pre-trained translation pair scoring system. We then generate "hard-negatives" following Guo et al. (2018) by using a pre-trained coarse translation-ranking model to determine translations that are close to a correct translation. As mentioned in the paper, the final constructed corpus contains around 600M en-fr pairs, 470M en-es pairs and 500M en-de pairs.

A.2 DOWNSTREAM TASK DESCRIPTIONS

Each of the English downstream tasks shown in Table 1 in the main body of the paper are described briefly below:

- **Movie Reviews (MR):** Multi-class prediction (scale of one to five stars) on movie review snippet data from Pang & Lee (2005).
- **Customer Reviews (CR):** Binary sentiment classification of sentences from customer reviews mined by Hu & Liu (2004).
- **Subjectivity (SUBJ):** Binary classification of sentences from movie reviews (Pang & Lee, 2004) as either objective or subjective.

- **Multi-Perspective Question Answering (MPQA):** Multi-class opinion polarity prediction on news data from Wiebe et al. (2005).
- **Text REtrieval Conference (TREC):** Multi-class classification of questions obtained from TREC by Li & Roth (2002).
- **Stanford Sentiment Treebank (SST):** Phrase-level binary sentiment classification of text from Socher et al. (2013).
- **Semantic Text Similarity (STS):** Semantic textual similarity of sentence pairs in the form of Pearson correlation with human judgments (Cer et al., 2017).

A.3 State-Of-The-Art STS System Descriptions

For the STS evaluations provided in the body of our paper, we compared to the state-of-the-art systems of Tian et al. (2017) and Wu et al. (2017). We refer the reader to Cer et al. (2017) for a detailed summary of the characteristics of these systems, but we provide a brief characterization below for convenience:

- **ECNU:** The ECNU model is an ensemble of random forest, gradient boosted decision trees, and deep learning models that leverages several manually constructed features such as n-gram overlap, edit distance, longest common prefix, common substrings, word alignments, and more.
- **BIT:** The BIT model relies mostly on a feature called information content (IC), which is based on the likelihood of occurrence of a concept (in this case, a word). Information content is then computed and combined for sentences through a hierarchical approach, which Wu et al. (2017) also augment with word embedding alignment to produce a final ensemble model.

A.4 Further Experiments

Given that the multi-task models detailed in the body of the paper are trained with an English NLI task but no non-English NLI tasks, we evaluate how much training without this English NLI task affects model performance on downstream English tasks. The performance of these no-NLI models are summarized under the "Cross-lingual Multitask Transformer No SNLI" section of Table 8. We find that training without SNLI leads to comparable or better performance on all English downstream tasks except for STS, where we find training with SNLI provides a significant bump in performance.

Additionally, the cross-lingual, multi-task models discussed in the body of our paper use a largely parallel task setup, where for each source-target language pair the monolingual target tasks are mirrored from the monolingual source tasks (the only exception being NLI, which is English-only). To test how useful this mirroring of monolingual tasks is, we train several models without target language monolingual tasks. We label these models as the "no non-English native task" models, and also summarize their downstream English task performance in Table 8. We note that removing the non-English native tasks actually leads to significant decreases in performance on TREC, SST, and STS. We find this quite interesting, as it provides some more empirical justification for the notion that cross-lingual training for a source-target language pair can actually improve monolingual source task performance.

Finally, we also evaluate the zero-shot performance of the no non-English native task models on Amazon review sentiment classification. As can be seen in Table 9, training without the non-English native tasks leads to lower performance on English and roughly the same zero-shot classification performance as training with only translation pair data.

Table 8: English task performance of different model configurations. Notably, removing non-English tasks in encoder training actually hurts performance on downstream English tasks.

Model	**MR**	**CR**	**SUBJ**	**MPQA**	**TREC**	**SST**	STS Benchmark (dev / test)
Cross-lingual Multitask Transformer SNLI + Native Tasks (report above)							
en-fr	77.9	82.9	95.5	89.3	95.3	84.0	0.803 / 0.763
en-es	80.1	85.9	94.6	86.5	96.2	85.2	0.809 / 0.770
en-de	78.8	84.0	95.9	87.6	96.1	85.0	0.802 / 0.764
Cross-lingual Multitask Transformer No SNLI							
en-fr	80.5	86.5	94.0	89.1	96.6	85.0	0.747 / 0.722
en-es	80.0	89.4	95.6	90.5	94.5	85.7	0.754 / 0.730
en-de	80.5	85.5	94.0	90.0	92.8	82.6	0.737 / 0.723
Cross-lingual Multitask Transformer No non-English native tasks							
en-fr	79.4	84.0	93.5	89.7	93.8	83.4	0.797 / 0.758
en-es	77.4	81.7	94.9	89.4	94.0	82.3	0.796 / 0.761
en-de	79.3	83.7	94.0	88.1	91.2	82.5	0.760 / 0.732
Translation-ranking Models							
en-fr	68.7	79.3	87.0	81.8	89.4	74.2	0.668 / 0.558
en-es	67.7	75.7	83.5	86.0	94.4	72.6	0.669 / 0.631
en-de	67.8	75.2	84.4	83.6	86.8	74.6	0.673 / 0.632

Table 9: Zero-shot sentiment classification accuracy(%) on target language Amazon review test data after training on only English Amazon review data.

Model	**en**	**fr**	**de**
Multi-task en-fr	87.4	82.3	–
Multi-task no non-English native tasks en-fr	85.6	65.4	–
Translation-ranking en-fr	74.4	66.3	–
Multi-task en-de	87.1	–	81.0
Multi-task No non-English native tasks en-fr	85.1	–	67.2
Translation-ranking en-de	73.8	–	67.0