Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model

Muthuraman Chidambaram[*] **Yinfei Yang**[*] **Daniel Cer**[*] **Steve Yuan,**
Yun-Hsuan Sung, Brian Strope, Ray Kurzweil
Google AI, Mountain View, CA, USA
{mutty, yinfeiy, cer}@google.com

Abstract

The scarcity of labeled training data across many languages is a significant roadblock for multilingual neural language processing. We approach the lack of in-language training data using sentence embeddings that map text written in different languages, but with similar meanings, to nearby embedding space representations. The representations are produced using a dual-encoder based model trained to maximize the representational similarity between sentence pairs drawn from parallel data. The representations are enhanced using multitask training and unsupervised monolingual corpora. The effectiveness of our multilingual sentence embeddings are assessed on a comprehensive collection of monolingual, cross-lingual, and zero-shot/few-shot learning tasks.

1 Introduction

Sentence embeddings are broadly useful for a diverse collection of downstream natural language processing tasks (Cer et al., 2018; Conneau et al., 2017; Kiros et al., 2015; Logeswaran and Lee, 2018; Subramanian et al., 2018). Sentence embeddings evaluated on downstream tasks in prior work have been trained on monolingual data, preventing them from being used for cross-lingual transfer learning. However, recent work on learning multilingual sentence embeddings has produced representations that capture semantic similarity even when sentences are written in different languages (Eriguchi et al., 2018; Guo et al., 2018; Schwenk and Douze, 2017; Singla et al., 2018). *We explore multi-task extensions of multilingual models for cross-lingual transfer learning.*

We present a novel approach for cross-lingual representation learning that combines methods for multi-task learning of monolingual sentence representations (Cer et al., 2018; Subramanian et al., 2018) with recent work on dual encoder methods for obtaining multilingual sentence representations for bi-text retrieval (Guo et al., 2018; Yang et al., 2019). By doing so, we learn representations that maintain strong performance on the original monolingual language tasks, while *simultaneously* obtaining good performance using zero-shot learning on the same task in another language. For a given language pair, we construct a multi-task training scheme using native source language tasks, native target language tasks, and a *bridging translation task* to encourage sentences with identical meanings, but written in different languages, to have similar embeddings.

We evaluate the learned representations on several monolingual and cross-lingual tasks, and provide a graph-based analysis of the learned representations. Multi-task training using additional monolingual tasks is found to improve performance over models that only make use of parallel data on both cross-lingual semantic textual similarity (STS) (Cer et al., 2017) and cross-lingual eigen-similarity (Søgaard et al., 2018). For European languages, the results show that the addition of monolingual data improves the embedding alignment of sentences and their translations. Further, we find that cross-lingual training with additional monolingual data leads to far better cross-lingual transfer learning performance.[1]

[*] equal contribution

[1] Models based on this work are available at https://tfhub.dev/ as: universal-sentence-encoder-xling/en-de, universal-sentence-encoder-xling/en-fr, and universal-sentence-encoder-xling/en-es. A large multilingual model is available as universal-sentence-encoder-xling/many.

Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 250–259
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

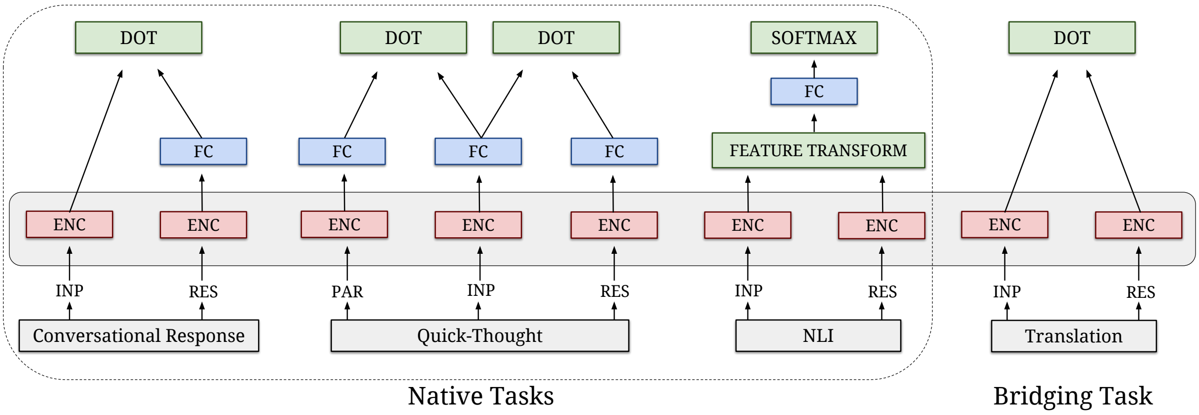

Figure 1: Multi-task dual-encoder model with native tasks and a bridging translation task. The terms PAR, INP, RES refer to parent, input, and response respectively. ENC refers to the shared encoder g, FC refers to fully connected layers, and DOT refers to dot product. Finally, FEATURE TRANSFORM refers to the feature vector used for natural language inference.

2 Multi-Task Dual-Encoder Model

The core of our approach is multi-task training over problems that can be modeled as ranking input-response pairs encoded via dual-encoders (Cer et al., 2018; Henderson et al., 2017; Yang et al., 2018). Cross-lingual representations are obtained by incorporating a *translation bridge task* (Gouws et al., 2015; Guo et al., 2018; Yang et al., 2019). For input-response ranking, 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 $s_j^R \in \mathcal{S}^R$. We model the conditional probability $P(s_i^R \mid s_i^I)$ as:

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

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

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

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

To obtain a single sentence encoding function g for use in downstream tasks, we share the first k layers of the input and response encoders and treat the final output of these shared layers as g. The shared encoders are used with the ranking formulation above to support conversational response ranking (Henderson et al., 2017), a modified version of quick-though (Logeswaran and Lee, 2018), and a supervised NLI task for representation learning similar to InferSent (Conneau et al., 2017). To learn cross-lingual representations, we incorporate translation ranking tasks using parallel corpora for the source-target pairs: English-French (en-fr), English-Spanish (en-es), English-German (en-de), and English-Chinese (en-zh).

The resulting model structure is illustrated in Figure 1. We note that the conversational response ranking task can be seen as a special case of Contrastive Predictive Coding (CPC) (van den Oord et al., 2018) that only makes predictions one step into the future.

2.1 Encoder Architecture

Word and Character Embeddings. Our sentence encoder makes use of word and character n-gram embeddings. Word embeddings are learned end-to-end.[2] Character n-gram embeddings are learned in a similar manner and are combined at the word-level by summing their representations and then passing the resulting vector to a single

[2] Using pre-trained embeddings, did not improve performance during preliminary experiments.

feedforward layer with $tanh$ activation. We average the word and character embeddings before providing them as input to g.

Transformer Encoder. The architecture of the shared encoder g consists of three stacked transformer sub-networks,[3] each containing the feedforward and multi-head attention sub-layers described in Vaswani et al. (2017). The transformer output is a variable-length sequence. We average encodings of all sequence positions in the final layer to obtain our sentence embeddings. 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.

2.2 Multi-task Training Setup

We employ four unique task types for each language pair in order to learn a function g that is capable of strong cross-lingual semantic matching and transfer learning performance for a source-target language pair while also maintaining monolingual task transfer performance. Specifically, we employ: *(i) conversational response prediction, (ii) quick thought, (iii) a natural language inference*, and *(iv) translation ranking* as the bridge task. For models trained on a single language pair (e.g., en-fr), six total tasks are used in training, as the first two tasks are mirrored across languages.[4]

Conversational Response Prediction. We model the conversational 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 $g(s_i^R)$ followed by two fully-connected feedforward layers of size 320 and 512 with $tanh$ activation. For the input representation, however, we simply let $g^I(s_i^I) = g(s_i^I)$.[5]

Quick Thought. We use a modified version of the Quick Thought task detailed by Logeswaran and 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 $g^R(s_i^R)$ by g followed by separate, fully-connected feedforward layers of size 320 and 512 and using $tanh$ activation.

Natural Language Inference (NLI). We also include an *English-only* natural language inference task (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. Following Conneau et al. (2017), the vectors u_1, u_2 are then used to construct a relation feature vector $(u_1, u_2, |u_1 - u_2|, u_1 * u_2)$, where $(\cdot)$ represents concatenation and $*$ represents element-wise multiplication. The relation vector is then fed into a single feedforward layer of size 512 followed by a softmax output layer 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 of 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 incorrect translations that are semantically similar to the correct translation for 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.

3 Experiments

3.1 Corpora

Training data is composed of Reddit, Wikipedia, Stanford Natural Language Inference (SNLI), and web mined translation pairs. For each of our datasets, we use 90% of the data for training, and the remaining 10% for development/validation.

3.2 Model Configuration

In all of our experiments, multi-task training is performed 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

[3] We tried up to six stacked transformers, but did not notice a significant difference beyond three.

[4] We note that our architecture can scale to models trained on > 2 languages. Preliminary experiments using more than two languages achieve promising results, but we consider fully evaluating models trained on larger collections of languages to be outside the scope of the current work.

[5] In early experiments, 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.

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
en-zh	76.1	83.4	93.0	86.4	**97.7**	81.4	0.791 / 0.770
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
en-zh	73.6	78.5	88.1	88.2	96.1	77.1	0.779 / 0.761
Prior Work							
CPC (van den Oord et al., 2018)	76.9	80.1	91.2	87.7	96.8	–	–
USE Trans. (Cer et al., 2018)	**81.4**	**87.4**	93.9	87.0	92.5	85.4	**0.814 / 0.782**
QT (Logeswaran and Lee, 2018)	82.4	86.0	94.8	**90.2**	92.4	**87.6**	–
InferSent (Conneau et al., 2017)	81.1	86.3	92.4	**90.2**	88.2	84.6	0.801 / 0.758
ST LN (Kiros et al., 2015)	79.4	83.1	93.7	89.3	–	–	–

Table 1: Performance on classification transfer tasks from SentEval (Conneau and Kiela, 2018).

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 treebank style 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 the development portion (random 10% slice) of our training sets. As an additional training heuristic, we multiply the gradient updates to the word and character embeddings by a factor of 100.[6] We found that using this embedding gradient multiplier alleviates vanishing gradients and greatly improves training.

We compare the proposed cross-lingual multi-task models, subsequently referred to simply as “multi-task", with baseline models that are trained using only the translation ranking task, referred to as “translation-ranking” models.

3.3 Model Performance on English Downstream Tasks

We first evaluate all of our cross-lingual models on several downstream English tasks taken from SentEval (Conneau and Kiela, 2018) to verify the impact of cross-lingual training. Evaluations are performed by training single hidden-layer feedforward networks on top of the 512-dimensional embeddings taken from the frozen models. 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 by a sizable amount.

We observe the en-zh translation-ranking models perform significantly better on the downstream tasks than the European language pair translation-ranking models. The en-zh models are possibly less capable of exploiting grammatical and other superficial similarities and are forced to rely on semantic representations. Exploring this further may present a promising direction for future research.

3.4 Cross-lingual Retrieval

We evaluate both the multi-task and translation-ranking models' efficacy in performing cross-lingual retrieval by using held-out translation pair data. Following Guo et al. (2018) and Henderson et al. (2017), we use precision at N (P@N) as our evaluation metric. Performance is scored by checking if a source sentence's target translation ranks[7] in the top N scored candidates when considering K other randomly selected target sentences. We set K to 999. Similar to Guo et al. (2018), we observe using a small value of K, such as $K = 99$ from Henderson et al. (2017), results

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

[7] Translation ranking scores are obtained by the dot product of source and target representations

Model	STS Benchmark (dev / test)				
	en	fr	es	de	zh
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**	–	–
Trans.-ranking en-es	0.669 / 0.631	–	0.622 / 0.611	–	–
Multi-task en-de	0.802 / 0.764	–	–	**0.768 / 0.722**	–
Trans.-ranking en-de	0.673 / 0.632	–	–	0.630 / 0.526	–
Multi-task en-zh	0.791 / **0.770**	–	–	–	0.730 / **0.705**
Trans.-ranking en-zh	0.779 / 0.761	–	–	–	**0.733** / 0.701

Table 2: Pearson's correlation coefficients on STS Benchmark (dev / test). The first column shows the results on the original STS Benchmark data in English. French, Spanish

in all metrics quickly obtaining $> 99\%$ P@1.[8]

The translation-ranking model is a strong baseline for identifying correct translations, with 95.4%, 87.5%, 97.5%, and 99.7% P@1 for en-fr, en-es, en-de, and en-zh retrieval tasks, respectively. The multi-task model performs almost identical with 95.1%, 88.8%, 97.8%, and 99.7% P@1, which provides empirical justification that it is possible to maintain cross-lingual embedding space alignment despite training on additional monolingual tasks for each individual language.[9] Both model types surprisingly achieve particularly strong ranking performance on en-zh. Similar to the task transfer experiments, this may be due to the en-zh models having an implicit inductive bias to rely more heavily on semantics rather than more superficial aspects of sentence pair similarity.

3.5 Multilingual STS

Cross-lingual representations are evaluated on semantic textual similarity (STS) in French, Spanish, German, and Chinese. To evaluate Spanish-Spanish (es-es) STS, we use data from track 3 of the SemEval-2017 STS shared task (Cer et al., 2017), containing 250 Spanish sentence pairs. We evaluate English-Spanish (en-es) STS using STS 2017 track 4(a),[10] which contains 250 English-Spanish sentence pairs.

Beyond English and Spanish, however, there are no standard STS datasets available for the other languages explored in this work. As such, we perform an additional evaluation on a translated version of the STS Benchmark (Cer et al., 2017) for French, Spanish, German, and Chinese. We use Google's translation system to translate the STS Benchmark sentences into each of these languages. We believe the results on the translated STS Benchmark evaluation sets are a reasonable indicator of multilingual semantic similarly performance, particularly 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 embeddings u, v for an STS sentence pair, and then score the sentence pair similarity based on the angular distance between the two embedding vectors, $-\arccos\left(\frac{uv}{||u||\,||v||}\right)$. Table 2 shows Pearson's r on the STS Benchmark for all models. The first column shows the trained model performance on the original English STS Benchmark. Columns 2 to 5 provide the performance on the remaining languages. Multi-task models perform better than the translation ranking models on our multilingual STS Benchmark evaluation sets. Table 3 provides the results from the en-es models on the SemEval-2017 STS *-es tracks. The multi-task models achieve 0.827 Pearson's r for the es-es task and 0.769 for the en-es task. As a point of reference, we also list the two best performing STS systems, ECNU (Tian et al., 2017) and BIT (Wu et al., 2017), as reported in Cer et al. (2017). Our results are very close to these state-of-the-art feature engineered and mixed systems.

[8] 999 is smaller than the 10+ million used by Guo et al. (2018), but it allows for good discrimination between models without requiring a heavier and slower evaluation framework

[9] We also experimented with P@3 and P@10, the results are identical.

[10] The en-es task is split into track 4(a) and track 4(b). We only use track 4(a) here. Track 4(b) contains sentence pairs from WMT with only one annotator for each pair. Previously reported numbers are particularly low for track 4(b), which may suggest either distributional or annotation differences between this track and other STS datasets.

Model	STS (SemEval 2017)	
	es-es	en-es
Multi-task	0.827	0.769
Trans.-ranking	0.642	0.587
ECNU	**0.856**	**0.813**
BIT	0.846	0.749

Table 3: Pearson's r on track 3 (es-es) and track 4(a) (en-es) of the SemEval-2017 STS shared task.

4 Zero-shot Classification

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

4.1 Multilingual NLI

We evaluate the zero-shot classification performance of our multi-task models on two multilingual natural language inference (NLI) tasks. However, prior to doing so, we first train a modified version[11] of our multi-task models that also includes training on the English Multi-genre NLI (MultiNLI) dataset of Williams et al. (2018) in addition to SNLI. We train with MultiNLI to be consistent with the baselines from prior work.

We make use of the professionally translated French and Spanish SNLI subset created by Agić and Schluter (2018) for an initial cross-lingual zero-shot evaluation of French and Spanish. We refer to these translated subsets as SNLI-X. There are 1,000 examples in the subset for each language. To evaluate, we feed the French and Spanish examples into the pre-trained English NLI sub-network of our multi-task models.

We additionally make use of the XNLI dataset of Conneau et al. (2018), which provides multilingual NLI evaluations for Spanish, French, German, Chinese and more. There are 5,000 examples in each XNLI test set, and zero-shot evaluation is once again done by feeding non-English examples into the pre-trained English NLI sub-network.

Table 4 lists the accuracy on the English SNLI test set as well as on SNLI-X and XNLI for all of our multi-task 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 SNLI-X is around 74% for both the en-fr and en-es models. The zero-shot accuracy on XNLI is around 65% for en-es, en-fr, and en-de, and around 63% for en-zh, thereby significantly outperforming the pretrained sentence encoding baselines (X-CBOW) described in Conneau et al. (2018). The X-CBOW baselines use fixed sentence encoders that are the result of averaging tuned multilingual word embeddings.

Row 4 of Table 4 shows the zero-shot French NLI performance of Eriguchi et al. (2018), which is a state-of-the-art zero-shot NLI classifier 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 Reviews

Zero-shot Learning. We also conduct a zero-shot evaluation based on the Amazon review data extracted by Prettenhofer and Stein (2010). Following Prettenhofer and Stein (2010), we preprocess the Amazon reviews and convert the data into a binary sentiment classification task by considering reviews with strictly more than three stars as positive and less than three stars as negative. Reviews contain a summary field and a text field, which we concatenate to produce a single input. Since our models are trained with sentence lengths clipped to 64, we only take the first 64 tokens from the concatenated text as the input. There are 6,000 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 hidden layers of size 512 with $tanh$ activation functions. 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 6,000 French and German Amazon review test examples. The results are summarized in Table 5. On the English test set, accuracy of the en-fr model is 87.4% with the en-de model achieving 87.1%. Both mod-

[11] Training with additional MultiNLI data did not significantly impact SNLI or downstream task performance.

Model	SNLI-X			XNLI				
	en	fr	es	en	fr	es	de	zh
Multi-task en-fr	84.2	**74.0**	–	**71.6**	**64.4**	–	–	–
Multi-task en-es	83.9	–	75.9	70.2	–	**65.2**	–	–
Multi-task en-de	84.1	–	–	71.5	–	–	**65.0**	–
Multi-task en-zh	83.7	–	–	69.2	–	–	–	**62.8**
NMT en-fr (Eriguchi et al., 2018)	**84.4**	73.9	–		–	–	–	–
XNLI-CBOW zero-shot (Conneau et al., 2018)	–	–	–	64.5	60.3	60.7	61.0	58.8
Non zero-shot baselines								
XNLI-BiLSTM-last (Conneau et al., 2018)	–	–	–	71.0	65.2	67.8	66.6	63.7
XNLI-BiLSTM-max (Conneau et al., 2018)	–	–	–	**73.7**	**67.7**	**68.7**	**67.7**	**65.8**

Table 4: Zero-shot classification accuracy (%) on SNLI-X and XNLI datasets. Cross-lingual transfer models are training on English only NLI data and then evaluated on French (fr), Spanish (es), German (de) and Chinese (zh) evaluation sets.

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	–

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

els achieve zero-shot accuracy on their respective non-English datasets that is above 80%. 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 French test set.[12]

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 trained on French and German reviews only. Table 6 provides the classification accuracy of the few-shot models, where the second row indicates 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 well compare to 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. The French model notably performs +0.9% better when being trained on a combination of the English and French reviews rather than on the French reviews alone.

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. We take N samples of size K from the language pair translation data and then encode these samples using the corresponding multi-task and translation-ranking models. We then compute pairwise distance matrices within each sampled set of encodings, and use these distance matrices to construct graph Laplacians.[13] We obtain the similarity $\Psi(S, T)$ between each model's source and target language embedding by comparing the eigenvalues of the source language graph Laplacians to the eigenvalues of the target language graph Laplacians:

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

Where $L_i^{(s)}$ and $L_i^{(t)}$ refer to the graph Laplacians of the source language and target language sentences obtained from the i^{th} sample of

[12] 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.

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

Target Language	Model		% available fr/de data					
			0%	10%	20%	40%	80%	100%
French	Few-shot	100% en + X% fr	82.3	**84.4**	**84.4**	**84.8**	**85.2**	**85.8**
	Monolingual	0% en + X% fr	–	79.2	80.0	82.7	84.3	84.9
German	Few-shot	100% en + X% de	81.0	**81.6**	**83.3**	**84.0**	**84.7**	**84.5**
	Monolingual	0% en + X% de	–	75.5	77.7	81.6	83.5	84.4

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.

source-target translation pairs. A smaller value of $\Psi(S, T)$ indicates higher eigen-similarity between 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 draw $N = 1,000$ samples instead of $N = 10$, as was done in Søgaard et al. (2018). We found $\Psi(S, T)$ has 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.

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 the European source-target language pairs, and observe the opposite for the English-Chinese models (en-zh). As a curious discrepancy, we believe further experiments looking at eigen-similarity across languages could yield interesting results and language groupings.

Eigen-similarity trends with better performance for the European language pair multi-task models on the cross-lingual transfer tasks. A potential direction for future work could be to introduce regularization penalties based on graph similarity during 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. Thus, another direction could be to further study differences in the difficulty of aligning different source-target language embeddings.

Model	en-fr	en-es	en-de	en-zh
multi-task	**0.592**	**0.526**	**0.761**	2.366
trans.-ranking	1.036	0.572	2.187	**0.393**

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

5.1 Discussion on Input Representations

Our early explorations using a combination of character n-gram embeddings and word embeddings vs. word embeddings alone as the model input representation suggest using word-embeddings only performs just slightly worse (one to two absolute percentage points) on the dev sets for the training tasks. The notable exception is the word-embedding only English-German models tend to perform much worse on the dev sets for the training tasks involving German. This is likely due to the prevalence of compound words in German and represents an interesting difference for future exploration.

We subsequently explored training versions of our cross-lingual models using a SentencePiece vocabulary (Kudo and Richardson, 2018), a set of largely sub-word tokens (characters and word chunks) that provide good coverage of an input dataset. Multilingual models for a single language pair (e.g., en-de) trained with SentencePiece performed similarly on the training dev sets to the models using character n-grams. However, when more languages are included in a single model (e.g., a single model that covers en, fr, de, es, and zh), SentencePiece tends to perform worse than using a combination of word and character n-gram embeddings. Within a larger joint model, SentencePiece is particularly problematic for languages like zh, which end up getting largely tokenized into individual characters.

6 Conclusion

Cross-lingual multi-task dual-encoder models are found to learn representations that achieve strong within language and cross-lingual transfer learning performance. By training English-French, English-Spanish, English-German, and English-

Chinese multi-task models, we achieve near-state-of-the-art or state-of-the-art performance on a variety of English tasks, while also being able to produce similar caliber results in zero-shot cross-lingual transfer learning tasks. Further, cross-lingual multi-task training is shown to improve performance on some downstream English tasks (TREC). We believe that there are many possibilities for future explorations of cross-lingual model training and that such models will be foundational as language processing systems are tasked with increasing amounts of multilingual data.

Acknowledgments

We thank the anonymous reviewers and our teammates from Descartes and other Google groups for their feedback and suggestions.

References

Željko Agić and Natalie Schluter. 2018. Baselines and test data for cross-lingual inference. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

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

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. 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. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel 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*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 748–756. JMLR.org.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Belgium, Brussels. Association for Computational Linguistics.

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

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 3294–3302, Cambridge, MA, USA. MIT Press.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Peter Prettenhofer and Benno Stein. 2010. Cross-Language Text Classification using Structural Correspondence Learning. In *48th Annual Meeting of the Association of Computational Linguistics (ACL 10)*, pages 1118–1127. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Karan Singla, Dogan Can, and Shrikanth Narayanan. 2018. 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)*, pages 214–220. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*.

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 2017. 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)*, pages 191–197, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Hao Wu, Heyan Huang, Ping Jian, Yuhang Guo, and Chao Su. 2017. 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)*, pages 77–84, Vancouver, Canada. Association for Computational Linguistics.

Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bidirectional dual encoder with additive margin softmax. In *Proceedings International Joint Conference on Artificial Intelligence (IJCAI)*.

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

X.-D. Zhang. 2011. The Laplacian eigenvalues of graphs: a survey. *CoRR*, abs/1111.2897.