Learning Sentence Representations for Cross-Lingual NLI with an Adversarial Objective

Asena D. Cengiz
New York University
Center for Data Science
asena@nyu.edu

Gauri Sarode
New York University
Tandon School of Engineering
gss383@nyu.edu

Sam Petter
New York University
Gallatin School of Individualized Studies
sjp591@nyu.edu

Abstract

Training reliable sentence encoders for natural language inference tasks requires large annotated data sets, which are readily available in English, but relatively sparse in other languages. The lack of non-English data limits the applicability of NLP systems in many real-world problems. However, recent studies demonstrate that it is possible to transfer knowledge from one language to another by building multilingual sentence encoders. Following this intuition, we train an LSTM sentence encoder on the Multi-genre Natural Language Inference (MultiNLI) corpus, transfer knowledge by aligning source and target encoders, and evaluate performance on Cross-Lingual Natural Language Inference (XNLI) tasks. We show that adding an adversarial loss to the alignment objective improves performance in 9 XNLI languages.

Our code is available on the linked Github repository.

1 Introduction

Natural Language Inference (Bowman et al., 2015), also known as Recognizing Textual Entailment ((RTE) Bar-Haim et al., 2006; Dagan et al., 2010), involves determining whether a sentence describing a situation, or *premise*, shares similar truth conditions, or *entails* another sentence called the *hypothesis*. Hypotheses with conflicting truth conditions are said to *contradict* the premise, and indeterminate relationships between the truth conditions of the two sentences are said to be *neutral*. Current machine learning systems designed to perform NLI tasks need to be trained on large amounts of premise-hypothesis pairs with entailment labels. Such large data sets in English have been published in recent years (Bowman et al., 2015; Williams et al., 2017; Khot et al., 2018; Zellers et al., 2018). However, many real-world applications of natural language models involve understanding and processing data in languages other than English. One method of transferring knowledge between languages for NLI tasks is to translate the training set to the target language or the development and test sets to the source language before inference time. Cross-Lingual Understanding (XLU), which is the process of learning to perform a task on a language and then transferring that knowledge to another language without reliance on translation, provides a simpler and more effective alternative (Conneau et al., 2018; Devlin et al., 2018; Lample and Conneau, 2019).

We focus on improving the encoder alignment procedure for cross-lingual understanding by incorporating an adversarial loss function to the alignment objective and we show that this method improves accuracy on XNLI tasks. The test performance of our model exceeds the XNLI baseline implementation in 9 languages. We start by training our encoder on the MultiNLI data set (Williams et al., 2017). Once the model is trained, we proceed to align source and target encoders using the parallel sentences from either Europarl (Koehn, 2005) or OpenSubtitles 2018 (Lison et al., 2018) corpora, depending on the target language. We jointly train the target encoder and a language discriminator to minimize the alignment objective (see Eq. 2). After aligning the target encoder to the source language space, we select the best-performing model using the XNLI development set and report test results on the XNLI test set for each language. The model exceeds the XNLI baseline in all languages except Bulgarian, Turkish, and Hindi. Table 1 presents the full set of results along with the baseline models. Since Swahili and Urdu are not included in the aligned multilingual word embedding

library (Joulin et al., 2018) that we use for lookup table initialization, we leave model development in these languages for future work.

2 Related Work

Natural Language Inference The two major approaches to Natural Language Inference involve either generating fixed-size sentence representations of the premise and the hypothesis and feeding a combination of those representations to a linear classifier to obtain class probabilities (Bowman et al., 2015; Talman et al., 2018; Chen et al., 2018) or utilizing attention between the premise and the hypothesis (Rocktäschel et al., 2015). As our goal is to build high-quality multilingual sentence encoders, we focus on the former. Natural Language Understanding research has been greatly assisted by the development of several large annotated corpora designed specifically for the NLI task (Bowman et al., 2015; Williams et al., 2017; Khot et al., 2018; Zellers et al., 2018).

MultiNLI The Multi-genre NLI Corpus consists of 392,702 training examples in five genres: fiction, government, slate, telephone, and travel (Williams et al., 2017). The development and test sets include both matched and unmatched genres. The current state-of-the-art on the MultiNLI dataset is held by Liu et al. (2019) with Multi-Task Deep Neural Networks (MTDNN).

XNLI The Cross-lingual Natural Language Inference (XNLI) corpus (Conneau et al., 2018) consists of development and test sets for MultiNLI corpus in 15 languages listed in Table 1. The data sets are generated by translating English premise-hypothesis pairs to each target language, rather than generating new pairs.

Multilingual Word Embeddings Multilingual word embeddings greatly assist cross-lingual understanding research. This is particularly salient in the field of translation as cross-lingual alignment of meaning operates as an extension of the distributional hypothesis (Faruqui and Dyer, 2014). Several methods of developing these embeddings have been used; Grave et al. (2018) train word vectors on Wikipedia and Common Crawl using Skip-gram and CBOW models (Mikolov et al., 2013) to obtain multilingual word vectors for 157 languages, and Joulin et al. (2018) retrieve aligned multilingual word vectors in 44 languages using Relaxed CSLS (RCSLS) as a similarity measure between continuous vectors. Xiao and Guo (2014) first create a dependency parse for a sentence both the target and original language and use this representation data to adjust the individual word embeddings. Many other cross-lingual alignment- and parsing-based methods have also proliferated in the field (Zou et al., 2013; Luong et al., 2015; Klementiev et al., 2012)

Sentence Representations One approach to mapping variable-length sentences into fixed-length vectors is to use Recurrent Neural Networks (Cho et al., 2014; Sutskever et al., 2014). Until recently, building sentence encoders that map text to vector space had been confined to English due to the lack of language understanding benchmarks in other, especially low-resource, languages. However, several sentence- or phrase-level encoders have been developed for the purposes of transfer learning, particularly those centered around sentence alignment (Wu et al., 2014; Cer et al., 2018; Conneau et al., 2018; Pham et al., 2015; Kaufmann, 2012; Wolk and Marasek, 2015; Zamani et al., 2016; Schwenk et al., 2017).

Adversarial Training In unsupervised machine translation, adversarial training is used for aligning sentence encoders (Lample et al., 2017; Conneau et al., 2017). Zellers et al. (2018) and Nie et al. (2018) use adversarial training as a means of weeding out overfitting in NLI models. A discriminator, which makes a binary prediction, is trained to predict the language of a given embedding correctly, while the encoder is trained to fool the discriminator. We transfer this approach to cross-lingual NLI setting by combining the adversarial loss (defined in Section [3.1.2]) with an l_2-norm-based alignment loss proposed by Conneau et al. (2018).

3 Experiments and Results

Our experiments follow three steps: i) Training a sentence encoder and a classifier on MultiNLI training set [3.1.1], ii) Aligning sentence encoders of English and the target languages using parallel corpora [3.1.2], and iii) Evaluating the cross-lingual NLI performance on XNLI development set of the target language [3.1.3]. Once we com-

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
XNLI Baselines from (Conneau et al., 2018)															
X-BiLSTM-last	71.0	65.2	67.8	66.6	66.3	65.7	63.7	64.2	62.7	65.6	62.7	63.7	62.8	54.1	56.4
X-BiLSTM-max	**73.7**	**67.7**	**68.7**	**67.7**	**68.9**	**67.9**	**65.4**	**64.2**	**64.8**	**66.4**	**64.1**	**65.8**	**64.1**	**55.7**	**58.4**
Baseline Results: *XNLI Multilingual Sentence Encoder (Our Implementation - Test Acc %)*															
X-BiLSTM-max	**70.1**	64.6	62.1	61.2	59.3	58.2	58.7	59.4	58.1	58.5	56.3	55.7	56.2	-	-
Model Results: *XNLI Multilingual Sentence Encoder + Discriminator (Test Acc %)*															
X-BiLSTM-max	**70.1**	65.0	64.5	61.5	59.7	58.1	59.0	58.3	58.6	60.0	57.1	56.6	55.9	-	-

Table 1: Cross-lingual natural language inference (XNLI) test accuracy for 13 languages.

plete all experiments, we choose the best model based on the development set performance and report test results for each XNLI language on Table 1.

3.1 Experiment Details

3.1.1 Training on MultiNLI

As a baseline for the multilingual sentence encoder, we implement the *X-BiLSTM-max* model from Conneau et al. (2018). We train a single-layer bidirectional LSTM as the sentence encoder, which is responsible for generating the fixed-size vector representations of sentences. The final feature vector is constructed by taking the element-wise maximum of the LSTM output across all states. The input vector of the linear classifier is:

$$[p, h, |p - h|, p \odot h] \tag{1}$$

where p and h are the sentence representations of the premise and the hypothesis, $|p - h|$ is the absolute value of their difference, and $\odot$ is element-wise multiplication. The 3-way log-softmax in the output layer computes log-probabilities for each class. Since the model output is the log-probabilities, we minimize Negative Log-Likelihood (NLL) Loss during training.

3.1.2 Aligning Sentence Encoders

In order to perform cross-lingual natural language inference, the target language encoder needs to generate sufficiently close sentence representations to the source encoder outputs. Since the goal is to trick the linear classifier to perceive the target encoder outputs as English - when only their representations are sufficiently close to English - we add a discriminator network on top of the alignment mechanism, to strengthen this effect; that is, we train the discriminator to differentiate between source and target languages while training the encoder to fool the discriminator. Our hypothesis is that this combined objective function will result in higher test accuracy for XNLI languages. The loss function used for this alignment by the baseline model is as follows:

$$\mathcal{L}_{XNLI}(x, y) = dist(x, y) - \lambda\,[dist(x_c, y) + dist(x, y_c)]$$

where $dist(x, y)$ is the $l2 - norm$, $||x - y||_2$, λ is the regularization coefficient, and x_c and y_c are contrastive sentence representations of the source and the target encoders. We obtain the contrastive samples via batch internal shuffling. We add an adversarial loss to the baseline loss function $\mathcal{L}_{XNLI}$ to compute our combined objective:

$$\mathcal{L}_{align} = \mathcal{L}_{XNLI} + \lambda_{adv}\mathcal{L}_{adv}(\theta_{enc}, \mathcal{W}|\theta_D) \tag{2}$$

where θ_{enc} and θ_D represent the encoder and the discriminator parameters, respectively, $\mathcal{W}$ is the sentence embedding and λ_{adv} is the weight of $\mathcal{L}_{adv}$. While the discriminator is trained to recognize the differences between languages, the setting in Eq. 2 ensures that the encoder is trained to fool the discriminator, because we define $\mathcal{L}_{adv}(\theta_{enc}, \mathcal{W}|\theta_D)$ as follows:

$$\begin{aligned}\mathcal{L}_{adv}(\theta_{enc}, \mathcal{W}|\theta_D) = &-\frac{1}{n}\sum_{i=1}^{n} logP_{\theta_D}(l_y = 1|\theta_{enc}, W_x) \\ &-\frac{1}{m}\sum_{i=1}^{m} logP_{\theta_D}(l_y = 0|\theta_{enc}, W_y)\end{aligned} \tag{3}$$

where l_x and l_y are the source and target language labels, 0 and 1; and $\mathcal{W}_x$ and $\mathcal{W}_y$ are the source and target sentence embeddings, respectively. The discriminator is trained with the following objective:

$$\begin{aligned}\mathcal{L}_D(\theta_D|\theta_{enc}, \mathcal{W}) = &-\frac{1}{n}\sum_{i=1}^{n} logP_{\theta_D}(l_y = 0|\theta_{enc}, W_x) \\ &-\frac{1}{m}\sum_{i=1}^{m} logP_{\theta_D}(l_y = 1|\theta_{enc}, W_y)\end{aligned} \tag{4}$$

Once trained on a combination of two monolingual corpora with language labels (0: source, 1: target), it should be able to recognize which

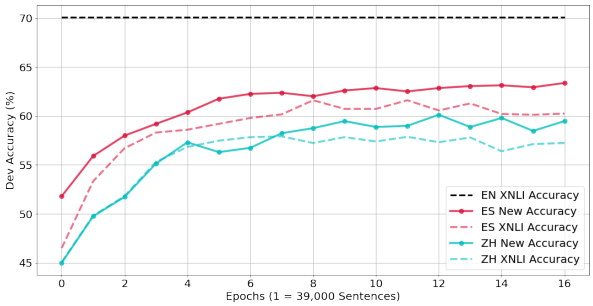

Figure 1: Figure 1: Evolution of Spanish and Chinese accuracies with the baseline model (-) and the new model (•).

Parameter	Value
λ in $\mathcal{L}_{align}$	0.25
Batch size	64
Encoder hidden size	512
# Discriminator layers	5
Dsc. leaky ReLU slope	0.28
Classifier nonlinearity	ReLU
Classifier hidden size	128
Optimizer	Adam

Table 2: Best multilingual encoder settings.

representation belongs to which language. By fixing the parameters of the English encoder, we allow the alignment loss to backpropagate only through the target encoder, ensuring that the target encoder outputs are aligned to the English embedding space.

Parallel Datasets for Alignment For French, German, Spanish, and Greek models we use Europarl Parallel Corpus (v7) that is collected from the proceedings of the European Parliament by Koehn (2005). For Turkish, Vietnamese, Bulgarian, Russian, Thai, Hindi, Chinese, and Arabic we use Open Subtitles 2018 corpus (Lison et al., 2018). We fix the maximum number of sentence pairs allowed to 2 million for each language, since above 2 million pairs we do not see any noteworthy improvement in development accuracy. We use the NLTK tokenizer for Arabic and Hindi, `jieba`[1] for Chinese, and the standard English tokenizer for the remaining languages.

3.1.3 Evaluation

After aligning the target sentence encoders to English embedding space we use the XNLI development sets select the best model and report the test accuracy for each language on Table 1.

Parameter Settings The parameter settings listed in Table 2 were optimized to best suit all languages used in the experiments. In addition to these, we fine tune the learning rate of the optimizer and dropout probability for all networks. We use the 250,000 most frequent words for each language, and no significant improvement in our development accuracy for increased vocabulary size was observed in any language. During model selection, we observed that the progress of the development set accuracy varied significantly among languages. For example, the Spanish encoder exceeds 60% dev-set accuracy with training on only 8% of the parallel sentences, whereas Arabic model reaches only around 50% accuracy with the same amount of data.

[1] https://github.com/fxsjy/jieba

3.2 Results

We saw an improvement the XNLI test set accuracy in the following XNLI languages: French (+0.4%), Spanish (+2.4%), German (+0.3%), Greek (+0.4%), Russian (+0.3%), Arabic (+0.5%), Vietnamese (+1.5%), Thai (+0.8%), and Chinese (+0.9%). Although the model performs exceptionally for those languages, for Bulgarian, Turkish, and Hindi, we don't see any improvement. We have included the full results on Table 1 along with the baseline model performance for each language.

4 Conclusion

The lack of annotated data required to train high-quality sentence encoders for NLI tasks presents a challenge for applications in languages other than English. Translation-based approaches reach state-of-the-art performance, however cross-lingual knowledge transfer provides a much simpler and more efficient alternative. By adding an adversarial loss to the sentence-level alignment objective, we show that the accuracy for cross-lingual Natural Language Inference can be improved. We observe that the models that already perform better (e.g. Spanish, French) than others improved even more, while models that were further away in the embedding space from the beginning struggled (see Figure 1).

5 Collaboration Statement

Tasks were distributed among group members in such a way that each member was able to participate in all aspects of the project. While each member guided the project when it came to their areas of expertise, work was evenly divided and equally understood.

References

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4. Venice.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

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.

Qian Chen, Zhen-Hua Ling, and Xiaodan Zhu. 2018. Enhancing sentence embedding with generalized pooling. *arXiv preprint arXiv:1806.09828*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

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

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

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

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984.

Max Kaufmann. 2012. JMaxAlign: A maximum entropy parallel sentence alignment tool. In *Proceedings of COLING 2012: Demonstration Papers*, pages 277–288, Mumbai, India. The COLING 2012 Organizing Committee.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India. The COLING 2012 Organizing Committee.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2018. Analyzing compositionality-sensitivity of NLI models. *CoRR*, abs/1811.07033.

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, Denver, Colorado. Association for Computational Linguistics.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

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

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.

Aarne Talman, Anssi Yli-Jyrä, and Jörg Tiedemann. 2018. Natural language inference with hierarchical bilstm max pooling architecture. *arXiv preprint arXiv:1808.08762*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Krzysztof Wolk and Krzysztof Marasek. 2015. A sentence meaning based alignment method for parallel text corpora preparation. *CoRR*, abs/1509.09093.

Haiyang Wu, Daxiang Dong, Xiaoguang Hu, Dianhai Yu, Wei He, Hua Wu, Haifeng Wang, and Ting Liu. 2014. Improve statistical machine translation with context-sensitive bilingual semantic embedding model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 142–146, Doha, Qatar. Association for Computational Linguistics.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129, Ann Arbor, Michigan. Association for Computational Linguistics.

Hamed Zamani, Heshaam Faili, and Azadeh Shakery. 2016. Sentence alignment using local and global information. *Comput. Speech Lang.*, 39(C):88–107.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. *CoRR*, abs/1808.05326.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, Washington, USA. Association for Computational Linguistics.