XNLI: Evaluating Cross-lingual Sentence Representations

Alexis Conneau, Ruty Rinott, Guillaume Lample, Holger Schwenk, Ves Stoyanov
Facebook AI
aconneau,ruty,glample,schwenk,ves@fb.com

Adina Williams, Samuel R. Bowman
New York University
adinawilliams,bowman@nyu.edu

Abstract

State-of-the-art natural language processing systems rely on supervision in the form of annotated data to learn competent models. These models are generally trained on data in a single language (usually English), and cannot be directly used beyond that language. Since collecting data in every language is not realistic, there has been a growing interest in cross-lingual language understanding (XLU) and low-resource cross-language transfer. In this work, we construct an evaluation set for XLU by extending the development and test sets of the Multi-Genre Natural Language Inference Corpus (MultiNLI) to 15 languages, including low-resource languages such as Swahili and Urdu. We hope that our dataset, dubbed XNLI, will catalyze research in cross-lingual sentence understanding by providing an informative standard evaluation task. In addition, we provide several baselines for multilingual sentence understanding, including two based on machine translation systems, and two that use parallel data to train aligned multilingual bag-of-words and LSTM encoders. We find that XNLI represents a practical and challenging evaluation suite, and that directly translating the test data yields the best performance among available baselines.

1 Introduction

Contemporary natural language processing systems typically rely on annotated data to learn how to perform a task (e.g., classification, sequence tagging, natural language inference). Most commonly the available training data is in a single language (e.g., English or Chinese) and the resulting system can perform the task only in the training language. In practice, however, systems used in major international products need to handle inputs in many languages. In these settings, it is nearly impossible to annotate data in all languages that a system might encounter during operation.

A scalable way to build multilingual systems is through cross-lingual language understanding (XLU), in which a system is trained primarily on data in one language and evaluated on data in others. While XLU shows promising results for tasks such as cross-lingual document classification (Klementiev et al., 2012; Schwenk and Li, 2018), there are very few, if any, XLU benchmarks for more difficult language understanding tasks like natural language inference. Large-scale natural language inference (NLI), also known as recognizing textual entailment (RTE), has emerged as a practical test bed for work on sentence understanding. In NLI, a system is tasked with reading two sentences and determining whether one entails the other, contradicts it, or neither (*neutral*). Recent crowdsourced annotation efforts have yielded datasets for NLI in English (Bowman et al., 2015; Williams et al., 2017) with nearly a million examples, and these have been widely used to evaluate neural network architectures and training strategies (Rocktäschel et al., 2016; Gong et al., 2018; Peters et al., 2018; Wang et al., 2018), as well as to train effective, reusable sentence representations (Conneau et al., 2017; Subramanian et al., 2018; Cer et al., 2018).

In this work, we introduce a benchmark that we call the Cross-lingual Natural Language Inference corpus, or XNLI, by extending these NLI corpora to 15 languages. XNLI consists of 7500 human-annotated development and test examples in NLI three-way classification format in English, French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu, making a total of 112,500 annotated pairs. These languages span several language families, and with the inclusion of Swahili and Urdu, include two lower-resource languages as well.

Because of its focus on development and test

Language	Premise / Hypothesis	Genre	Label
English	You don't have to stay there. You can leave.	Face-To-Face	Entailment
French	La figure 4 montre la courbe d'offre des services de partage de travaux. Les services de partage de travaux ont une offre variable.	Government	Entailment
Spanish	Y se estremeció con el recuerdo. El pensamiento sobre el acontecimiento hizo su estremecimiento.	Fiction	Entailment
German	Während der Depression war es die ärmste Gegend, kurz vor dem Hungertod. Die Weltwirtschaftskrise dauerte mehr als zehn Jahre an.	Travel	Neutral
Swahili	Ni silaha ya plastiki ya moja kwa moja inayopiga risasi. Inadumu zaidi kuliko silaha ya chuma.	Telephone	Neutral
Russian	И мы занимаемся этим уже на протяжении 85 лет. Мы только начали этим заниматься.	Letters	Contradiction
Chinese	让我告诉你，美国人最终如何看待你作为独立顾问的表现。 美国人完全不知道您是独立律师。	Slate	Contradiction
Arabic	تحتاج الوكالات لأن تكون قادرة على قياس مستويات النجاح. لا يمكنأللوكاات أ اتعرف ما إذا كانت ناجحة أم لا	Nine-Eleven	Contradiction

Table 1: Examples (premise and hypothesis) from various languages and genres from the XNLI corpus.

data, this corpus is designed to evaluate *cross-lingual* sentence understanding, where models have to be trained in one language and tested in different ones.

We evaluate several approaches to cross-lingual learning of natural language inference that leverage parallel data from publicly available corpora at training time. We show that parallel data can help align sentence encoders in multiple languages such that a classifier trained with English NLI data can correctly classify pairs of sentences in other languages. While outperformed by our machine translation baselines, we show that this alignment mechanism gives very competitive results.

A second practical use of XNLI is the evaluation of pretrained general-purpose language-universal sentence encoders. We hope that this benchmark will help the research community build multilingual text embedding spaces. Such embeddings spaces will facilitate the creation of multilingual systems that can transfer across languages with little or no extra supervision.

The paper is organized as follows: We next survey the related literature on cross-lingual language understanding. We then describe our data collection methods and the resulting corpus in Section 3. We describe our baselines in Section 4, and finally present and discuss results in Section 5.

2 Related Work

Multilingual Word Embeddings Much of the work on multilinguality in language understanding has been at the word level. Several approaches have been proposed to learn cross-lingual word representations, i.e., word representations where translations are close in the embedding space. Many of these methods require some form of supervision (typically in the form of a small bilingual lexicon) to align two sets of source and target embeddings to the same space (Mikolov et al., 2013a; Kociský et al., 2014; Faruqui and Dyer, 2014; Ammar et al., 2016). More recent studies have showed that cross-lingual word embeddings can be generated with no supervision whatsoever (Artetxe et al., 2017; Conneau et al., 2018).

Sentence Representation Learning Many approaches have been proposed to extend word embeddings to sentence or paragraph representations (Le and Mikolov, 2014; Wieting et al., 2016; Arora et al., 2017). The most straightforward way to generate sentence embeddings is to consider an average or weighted average of word representations, usually referred to as continuous bag-of-words (CBOW). Although naïve, this method often provides a strong baseline. More sophisticated approaches—such as the unsupervised SkipThought model of Kiros et al. (2015) that extends the skip-gram model of Mikolov et al. (2013b) to the sentence level—have been pro-

posed to capture syntactic and semantic dependencies inside sentence representations. While these fixed-size sentence embedding methods have been outperformed by their supervised counterparts (Conneau et al., 2017; Subramanian et al., 2018), some recent developments have shown that pretrained language models can also transfer very well, either when the hidden states of the model are used as contextualized word vectors (Peters et al., 2018), or when the full model is fine-tuned on transfer tasks (Radford et al.; Howard and Ruder, 2018).

Multilingual Sentence Representations There has been some effort on developing multilingual sentence embeddings. For example, Chandar et al. (2013) train bilingual autoencoders with the objective of minimizing reconstruction error between two languages. Schwenk et al. (2017) and España-Bonet et al. (2017) jointly train a sequence-to-sequence MT system on multiple languages to learn a shared multilingual sentence embedding space. Hermann and Blunsom (2014) propose a compositional vector model involving unigrams and bigrams to learn document level representations. Pham et al. (2015) directly train embedding representations for sentences with no attempt at compositionality. Zhou et al. (2016) learn bilingual document representations by minimizing the Euclidean distance between document representations and their translations.

Cross-lingual Evaluation Benchmarks The lack of evaluation benchmark has hindered the development of such multilingual representations. Most previous approaches use the Reuters cross-lingual document classification corpus Klementiev et al. (2012) for evaluation. However, the classification in this corpus is done at document level, and, as there are many ways to aggregate sentence embeddings, the comparison between different sentence embeddings is difficult. Moreover, the distribution of classes in the Reuters corpus is highly unbalanced, and the dataset does not provide a development set in the target language, further complicating experimental comparisons.

In addition to the Reuters corpus, Cer et al. (2017) propose sentence-level multilingual training and evaluation datasets for semantic textual similarity in four languages. There have also been efforts to build multilingual RTE datasets, either through translating English data (Mehdad et al., 2011), or annotating sentences from a parallel corpora (Negri et al., 2011). More recently, Agić and Schluter (2018) provide a corpus, that is very complementary to our work, of human translations for 1332 pairs of the SNLI data into Arabic, French, Russian, and Spanish. Among all these benchmarks, XNLI is the first large-scale corpus for evaluating sentence-level representations on that many languages.

In practice, cross-lingual sentence understanding goes beyond translation. For instance, Mohammad et al. (2016) analyze the differences in human sentiment annotations of Arabic sentences and their English translations, and conclude that most of them come from cultural differences. Similarly, Smith et al. (2016) show that most of the degradation in performance when applying a classification model trained in English to Spanish data translated to English is due to cultural differences. One of the limitations of the XNLI corpus is that it does not capture these differences, since it was obtained by translation. We see the XNLI evaluation as a necessary step for multilingual NLP before tackling the even more complex problem of domain-adaptation that occurs when handling this the change in style from one language to another.

3 The XNLI Corpus

Because the test portion of the Multi-Genre NLI data was kept private, the Cross-lingual NLI Corpus (XNLI) is based on new English NLI data. To collect the core English portion, we follow precisely the same crowdsourcing-based procedure used for the existing Multi-Genre NLI corpus, and collect and validate 750 new examples from each of the ten text sources used in that corpus for a total of 7500 examples. With that portion in place, we create the full XNLI corpus by employing professional translators to translate it into our ten target languages. This section describes this process and the resulting corpus.

Translating, rather than generating new hypothesis sentences in each language separately, has multiple advantages. First, it ensures that the data distributions are maximally similar across languages. As speakers of different languages may have slightly different intuitions about how to fill in the supplied prompt, this allows us to avoid adding this unwanted degree of freedom. Second, it allows us to use the same trusted pool of workers as was used prior NLI crowdsourcing efforts,

without the need for training a new pool of workers in each language. Third, for any premise, this process allows us to have a corresponding hypothesis in any language. XNLI can thus potentially be used to evaluate whether an Arabic or Urdu premise is entailed with a Bulgarian or French hypothesis etc. This results in more than 1.5M combinations of hypothesis and premises. Note that we do not consider that use case in this work.

This translation approach carries with it the risk that the semantic relations between the two sentences in each pair might not be reliably preserved in translation, as Mohammad et al. (2016) observed for sentiment. We investigate this potential issue in our corpus and find that, while it does occur, it only concerns a negligible number of sentences (see Section 3.2).

3.1 Data Collection

The English Corpus Our collection procedure for the English portion of the XNLI corpus follows the same procedure as the MultiNLI corpus. We sample 250 sentences from each of the ten sources that were used in that corpus, ensuring that none of those selected sentences overlap with the distributed corpus. Nine of the ten text sources are drawn from the second release of the Open American National Corpus[1]: *Face-To-Face*, *Telephone*, *Government*, *9/11*, *Letters*, *Oxford University Press* (*OUP*), *Slate*, *Verbatim*, and *Government*. The tenth, *Fiction*, is drawn from the novel *Captain Blood* (Sabatini, 1922). We refer the reader to Williams et al. (2017) for more details on each genre.

Given these sentences, we ask the same MultiNLI worker pool from a crowdsourcing platform to produce three hypotheses for each premise, one for each possible label.

We present premise sentences to workers using the same templates as were used in MultiNLI. We also follow that work in pursuing a second validation phase of data collection in which each pair of sentences is relabeled by four other workers. For each validated sentence pair, we assign a gold label representing a majority vote between the initial label assigned to the pair by the original annotator, and the four additional labels assigned by validation annotators. We obtained a three-vote consensus for 93% of the data. In our experiments, we kept the 7% additional ones, but we mark these ones with a special label '-'.

Translating the Corpus Finally, we hire translators to translate the resulting sentences into 15 languages using the *One Hour Translation* platform. We translate the premises and hypotheses separately, to ensure that no context is added to the hypothesis that was not there originally, and simply copy the labels from the English source text. Some development examples are shown in Table 1.

3.2 The Resulting Corpus

One main concern in studying the resulting corpus is to determine whether the gold label for some of the sentence pairs changes as a result of information added or removed in the translation process.

Investigating the data manually, we find an example in the Chinese translation where an entailment relation becomes a contradictory relation, while the entailment is preserved in other languages. Specifically, the term *upright* which was used in English as entailment of *standing*, was translated into Chinese as *sitting upright* thus creating a contradiction. However, the difficulty of finding such an example in the data suggests its rarity.

To quantify this observation, we recruit two bilingual annotators to re-annotate 100 examples each in both English and French following our standard validation procedure. The examples are drawn from two non-overlapping random subsets of the development data to prevent the annotators from seeing the source English text for any translated text they annotate. With no training or burn-in period, these annotators recover the English consensus label 85% of the time on the original English data and 83% of the time on the translated French, suggesting that the overall semantic relationship between the two languages has been preserved. As most sentences are relatively easy to translate, in particular the hypotheses generated by the workers, there seems to be little ambiguity added by the translator.

More broadly, we find that the resulting corpus has similar properties to the MultiNLI corpus. For all languages, on average, the premises are twice as long as the hypotheses (See Table 2). The top hypothesis words indicative of the class label – scored using the mutual information between each word and class in the corpus – are similar across languages, and overlap those of the MultiNLI cor-

[1] http://www.anc.org/

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
Premise	21.7	24.1	22.1	21.1	21.0	20.9	19.6	16.8	20.7	27.6	22.1	21.8	23.2	18.7	24.1
Hypothesis	10.7	12.4	10.9	10.8	10.6	10.4	9.7	8.4	10.2	13.5	10.4	10.8	11.9	9.0	12.3

Table 2: Average number of tokens per sentence in the XNLI corpus for each language.

pus (Gururangan et al., 2018). For example, a translation of at least one of the words *no*, *not* or *never* is among the top two cues for contradiction in all languages.

As in the original MultiNLI corpus, we expect that cues like these ('artifacts', in Gururangan's terms, also observed by Poliak et al., 2018; Tsuchiya, 2018) allow a baseline system to achieve better-than-random accuracy with access only to the premise sentences. We accept this as an unavoidable property of the NLI task over naturalistic sentence pairs, and see no reason to expect that this baseline would achieve better accuracy than the relatively poor 53% seen in Gururangan et al. (2018).

The current version of the corpus is freely available[2][3] for typical machine learning uses, and may be modified and redistributed. The current license is CC-BY-NC. The majority of the corpus sentences are released under the OANC's license which allows all content to be freely used, modified, and shared under permissive terms. The data in the *Fiction* genre from *Captain Blood* are in the public domain in the United States (but may be licensed differently elsewhere).

4 Cross-Lingual NLI

In this section we present results with XLU systems that can serve as baselines for future work.

4.1 Translation-Based Approaches

The most straightforward techniques for XLU rely on translation systems. There are two natural ways to use a translation system: TRANSLATE TRAIN, where the training data is translated into each target language to provide data to train each classifier, and TRANSLATE TEST, where a translation system is used at test time to translate input sentences to the training language. These two methods provide strong baselines, but both present practical challenges. The former requires training and maintaining as many classifiers as there are languages, while the latter relies on computationally-intensive translation at test time. Both approaches are limited by the quality of the translation system, which itself varies with the quantity of available training data and the similarity of the language pair involved.

4.2 Multilingual Sentence Encoders

An alternative to translation is to rely on language-universal embeddings of text and build multilingual classifiers on top of these representations. If an encoder produces an embedding of an English sentence close to the embedding of its translation in another language, then a classifier learned on top of English sentence embeddings will be able to classify sentences from different languages without needing a translation system at inference time.

We evaluate two types of cross-lingual sentence encoders: (i) pretrained universal multilingual sentence embeddings based on the average of word embeddings (X-CBOW), (ii) bidirectional-LSTM (BiLSTM) (Hochreiter and Schmidhuber, 1997) sentence encoders trained on the MultiNLI training data (X-BILSTM). The former evaluates transfer learning while the latter evaluates NLI-specific encoders trained on in-domain data. Both approaches use the same alignment loss for aligning sentence embedding spaces from multiple languages which is present below. We consider two ways of extracting feature vectors from the BiLSTM: either using the initial and final hidden states (Sutskever et al., 2014), or using the element-wise max over all states (Collobert and Weston, 2008).

The first approach is commonly used as a strong baseline for monolingual sentence embeddings (Arora et al., 2017; Conneau and Kiela, 2018; Gouews et al., 2014). Concretely, we consider the English fastText word embedding space as being fixed, and fine-tune embeddings in other languages so that the average of the word vectors in a sentence is close to the average of the word vectors in its English translation. The second ap-

[2] http://nyu.edu/projects/bowman/xnli/
[3] http://github/facebookresearch/xnli

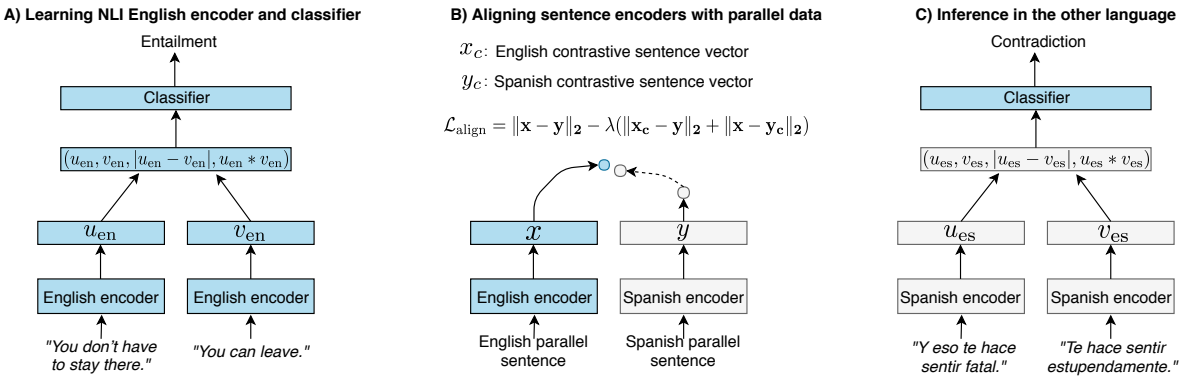

Figure 1: **Illustration of language adaptation by sentence embeddings alignment.** A) The English encoder and classifier in blue are learned on English (*in-domain*) NLI data. The encoder can also be pretrained (*transfer learning*). B) The Spanish encoder in gray is trained to mimic the English encoder using parallel data. C) After alignment of the encoders, the classifier can make predictions for Spanish.

proach consists in learning an English sentence encoder on the MultiNLI training data along with an encoder on the target language, with the objective that the representations of two translations are nearby in the embedding space. In both approaches, an English encoder is fixed, and we train target language encoders to match the output of this encoder. This allows us to build sentence representations that belong to the same space. Joint training of encoders and parameter sharing are also promising directions to improve and simplify the alignment of sentence embedding spaces. We leave this for future work.

In all experiments, we consider encoders that output a vector of fixed size as a sentence representation. While previous work shows that performance on the NLI task can be improved by using cross-sentence attention between the premise and hypothesis (Rocktäschel et al., 2016; Gong et al., 2018), we focus on methods with fixed-size sentence embeddings.

4.2.1 Aligning Word Embeddings

Multilingual word embeddings are an efficient way to transfer knowledge from one language to another. For instance, Zhang et al. (2016) show that cross-lingual embeddings can be used to extend an English part-of-speech tagger to the cross-lingual setting, and Xiao and Guo (2014) achieve similar results in dependency parsing. Cross-lingual embeddings also provide an efficient mechanism to bootstrap neural machine translation (NMT) systems for low-resource language pairs, which is critical in the case of unsupervised machine translation (Lample et al., 2018; Artetxe et al., 2018). In that case, the use cross-lingual embeddings directly helps the alignment of sentence-level encoders. Cross-lingual embeddings can be generated efficiently using a very small amount of supervision. By using a small parallel dictionary with $n = 5000$ word pairs, it is possible to learn a linear mapping to minimize

$$W^\star = \underset{W \in O_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_{\mathrm{F}} = UV^T,$$

where d is the dimension of the embeddings, and X and Y are two matrices of shape (d, n) that correspond to the aligned word embeddings that appear in the parallel dictionary, $O_d(\mathbb{R})$ is the group of orthogonal matrices of dimension d, and U and V are obtained from the singular value decomposition (SVD) of YX^T: $U\Sigma V^T = \mathrm{SVD}(YX^T)$. Xing et al. (2015) show that enforcing the orthogonality constraint on the linear mapping leads to better results on the word translation task.

In this paper, we pretrain our embeddings using the common-crawl word embeddings (Grave et al., 2018) aligned with the MUSE library of Conneau et al. (2018).

4.2.2 Universal Multilingual Sentence Embeddings

Most of the successful recent approaches for learning universal sentence representations have relied on English (Kiros et al., 2015; Arora et al., 2017; Conneau et al., 2017; Subramanian et al., 2018; Cer et al., 2018). While notable recent approaches have considered building a shared sentence encoder for multiple languages using publicly available parallel corpora (Johnson et al.,

	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
XX-En BLEU	41.2	45.8	39.3	42.1	38.7	27.1	29.9	35.2	23.6	22.6	24.6	27.3	21.3	24.4
En-XX BLEU	49.3	48.5	38.8	42.4	34.2	24.9	21.9	15.8	39.9	21.4	23.2	37.5	24.6	24.1
Word translation P@1	73.7	73.9	65.9	61.1	61.9	60.6	55.0	51.9	35.8	25.4	48.6	48.2	-	-

Table 3: BLEU scores of our translation models (XX-En) P@1 for multilingual word embeddings.

2016; Schwenk et al., 2017; España-Bonet et al., 2017), the lack of a large-scale, sentence-level semantic evaluation has limited their adoption by the community. In particular, these methods do not cover the scale of languages considered in XNLI, and are limited to high-resource languages. As a baseline for the evaluation of pretrained multilingual sentence representations in the 15 languages of XNLI, we consider state-of-the-art common-crawl embeddings with a CBOW encoder. Our approach, dubbed X-CBOW, consists in fixing the English pretrained word embeddings, and fine-tuning the target (e.g., French) word embeddings so that the CBOW representations of two translations are close in embedding space. In that case, we consider our multilingual sentence embeddings as being pretrained and only learn a classifier on top of them to evaluate their quality, similar to so-called "transfer" tasks in (Kiros et al., 2015; Conneau et al., 2017) but in the multilingual setting.

4.2.3 Aligning Sentence Embeddings

Training for similarity of source and target sentences in an embedding space is conceptually and computationally simpler than generating a translation in the target language from a source sentence. We propose a method for training for cross-lingual similarity and evaluate approaches based on the simpler task of aligning sentence representations. Under our objective, the embeddings of two parallel sentences need not be identical, but only close enough in the embedding space that the decision boundary of the English classifier captures the similarity.

We propose a simple alignment loss function to align the embedding spaces of two different languages. Specifically, we train an English encoder on NLI, and train a target encoder by minimizing the loss:

$$\mathcal{L}_{\text{align}}(x, y) = dist(x, y) - \lambda(dist(x_c, y) + dist(x, y_c))$$

where (x, y) corresponds to the source and target sentence embeddings, (x_c, y_c) is a contrastive term (i.e. negative sampling), λ controls the weight of the negative examples in the loss. For the distance measure, we use the L2 norm $dist(x, y) = \|x - y\|_2$. A ranking loss (Weston et al., 2011) of the form

$$\mathcal{L}_{\text{rank}}(x, y) = \max(0, \alpha - dist(x, y_c) + dist(x, y)) + \max(0, \alpha - dist(x_c, y) + dist(x, y))$$

that pushes the sentence embeddings of a translation pair to be closer than the ones of negative pairs leads to very poor results in this particular case. As opposed to $\mathcal{L}_{\text{align}}$, $\mathcal{L}_{\text{rank}}$ does not force the embeddings of sentence pairs to be close enough so that the shared classifier can understand that these sentences have the same meaning.

We use $\mathcal{L}_{\text{align}}$ in the cross-lingual embeddings baselines X-CBOW, X-BILSTM-LAST and X-BILSTM-MAX. For X-CBOW, the encoder is pretrained and not fine-tuned on NLI (transfer-learning), while the English X-BiLSTMs are trained on the MultiNLI training set (in-domain). For the three methods, the English encoder and classifier are then fixed. Each of the 14 other languages have their own encoders with same architecture. These encoders are trained to "copy" the English encoder using the $\mathcal{L}_{\text{align}}$ loss and the parallel data described in section 5.2. Our sentence embedding alignment approach is illustrated in Figure 1.

We only back-propagate through the target encoder when optimizing $\mathcal{L}_{\text{align}}$ such that all 14 encoders live in the same English embedding space. In these experiments, we initialize lookup tables of the LSTMs with pretrained cross-lingual embeddings discussed in Section 4.2.1.

5 Experiments and Results

5.1 Training details

We use internal translation systems to translate data between English and the 10 other languages. For TRANSLATE TEST (see Table 4), we translate each test set into English, while for the TRANSLATE TRAIN, we translate the English training

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
*Machine translation baselines (*TRANSLATE TRAIN*)*															
BiLSTM-last	71.0	66.7	67.0	65.7	65.3	65.6	65.1	61.9	63.9	63.1	61.3	65.7	61.3	55.2	55.2
BiLSTM-max	**73.7**	68.3	68.8	66.5	66.4	67.4	66.5	64.5	65.8	66.0	62.8	67.0	62.1	58.2	56.6
*Machine translation baselines (*TRANSLATE TEST*)*															
BiLSTM-last	71.0	68.3	68.7	66.9	67.3	68.1	66.2	64.9	65.8	64.3	63.2	66.5	61.8	60.1	58.1
BiLSTM-max	**73.7**	**70.4**	**70.7**	**68.7**	**69.1**	**70.4**	**67.8**	**66.3**	**66.8**	**66.5**	**64.4**	**68.3**	**64.2**	**61.8**	**59.3**
Evaluation of XNLI multilingual sentence encoders (in-domain)															
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
Evaluation of pretrained multilingual sentence encoders (transfer learning)															
X-CBOW	64.5	60.3	60.7	61.0	60.5	60.4	57.8	58.7	57.5	58.8	56.9	58.8	56.3	50.4	52.2

Table 4: Cross-lingual natural language inference (XNLI) test accuracy for the 15 languages.

data of MultiNLI[4]. To give an idea of the translation quality, we give BLEU scores of the automatic translation from the foreign language into English of the XNLI test set in Table 3. We use the MOSES tokenizer for most languages, falling back on the default English tokenizer when necessary. We use the Stanford segmenter for Chinese (Chang et al., 2008), and the *pythainlp* package for Thai.

We use pretrained 300D aligned word embeddings for both X-CBOW and X-BILSTM and only consider the most 500,000 frequent words in the dictionary, which generally covers more than 98% of the words found in XNLI corpora. We set the number of hidden units of the BiLSTMs to 512, and use the Adam optimizer (Kingma and Ba, 2014) with default parameters. As in (Conneau et al., 2017), the classifier receives a vector $[u, v, |u - v|, u * v]$, where u and v are the embeddings of the premise and hypothesis provided by the shared encoder, and $*$ corresponds to the element-wise multiplication (see Figure 1). For the alignment loss, setting λ to 0.25 worked best in our experiments, and we found that the tradeoff between the importance of the positive and the negative pairs was particularly important (see Table 5). We sample negatives randomly. When fitting the target BiLSTM encoder to the English encoder, we fine-tune the lookup table associated to the target encoder, but keep the source word embeddings fixed. The classifier is a feed-forward neural network with one hidden layer of 128 hidden units, regularized with dropout (Srivastava et al., 2014) at a rate of 0.1. For X-BiLSTMs, we perform model selection on the XNLI validation set in each target language. For X-CBOW, we keep a validation set of parallel sentences to evaluate our alignment loss. The alignment loss requires a parallel dataset of sentences for each pair of languages, which we describe next.

5.2 Parallel Datasets

We use publicly available parallel datasets to learn the alignment between English and target encoders. For French, Spanish, Russian, Arabic and Chinese, we use the United Nation corpora (Ziemski et al., 2016), for German, Greek and Bulgarian, the Europarl corpora (Koehn, 2005), for Turkish, Vietnamese and Thai, the OpenSubtitles 2018 corpus (Tiedemann, 2012), and for Hindi, the IIT Bombay corpus (Anoop et al., 2018). For all the above language pairs, we were able to gather more than 500,000 parallel sentences, and we set the maximum number of parallel sentences to 2 million. For the lower-resource languages Urdu and Swahili, the number of parallel sentences is an order of magnitude smaller than for the other languages we consider. For Urdu, we used the Bible and Quran transcriptions (Tiedemann, 2012), the OpenSubtitles 2016 and 2018 corpora (Tiedemann, 2012) and LDC2010T21, LDC2010T23 LDC corpora, and obtained a total of 64k parallel sentences. For Swahili, we were only able to gather 42k sentences using the Global Voices corpus and the Tanzil Quran transcription corpus[5].

[4]To allow replication of results, we share the MT translations of XNLI training and test sets.

[5]http://opus.nlpl.eu/

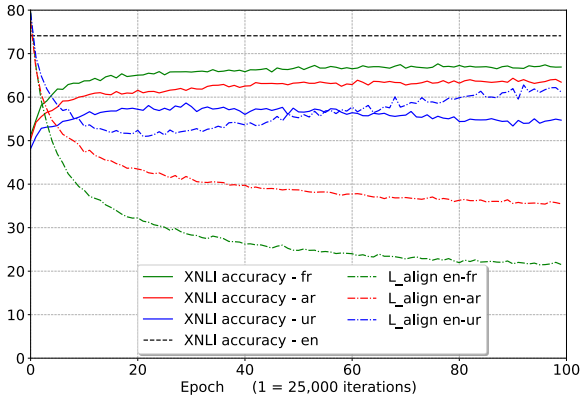

Figure 2: Evolution along training of alignment losses and X-BILSTM XNLI French (fr), Arabic (ar) and Urdu (ur) accuracies. Observe the correlation between $\mathcal{L}_{\text{align}}$ and accuracy.

5.3 Analysis

Comparing in-language performance in Table 4, we observe that, when using BiLSTMs, results are consistently better when we take the dimension-wise maximum over all hidden states (BiLSTM-max) compared to taking the last hidden state (BiLSTM-last). Unsuprisingly, BiLSTM results are better than the pretrained CBOW approach for all languages. As in Bowman et al. (2015), we also observe the superiority of BiLSTM encoders over CBOW, even when fine-tuning the word embeddings of the latter on the MultiNLI training set, thereby again confirming that the NLI task requires more than just word information. Both of these findings confirm previously published results (Conneau et al., 2017).

Table 4 shows that translation offers a strong baseline for XLU. Within translation, TRANSLATE TEST appears to perform consistently better than TRANSLATE TRAIN for all languages. The best cross-lingual results in our evaluation are obtained by the TRANSLATE TEST approach for all cross-lingual directions. Within the translation approaches, as expected, we observe that cross-lingual performance depends on the quality of the translation system. In fact, translation-based results are very well-correlated with the BLEU scores for the translation systems; XNLI performance for three of the four languages with the best translation systems (comparing absolute BLEU, Table 3) is above 70%. This performance is still about three points below the English NLI performance of 73.7%. This slight drop in performance may be related to translation error, changes in style, or artifacts introduced by the machine translation systems that result in discrepancies between the training and test data.

For cross-lingual performance, we observe a healthy gap between the English results and the results obtained on other languages. For instance, for French, we obtain 67.7% accuracy when classifying French pairs using our English classifier and multilingual sentence encoder. When using our alignment process, our method is competitive with the TRANSLATE TRAIN baseline, suggesting that it might be possible to encode similarity between languages directly in the embedding spaces generated by the encoders. However, these methods are still below the other machine translation baseline TRANSLATE TEST, which significantly outperforms the multilingual sentence encoder approach by up to 6% (Swahili). These production systems have been trained on much larger training data than the ones used for the alignment loss (section 5.2), which can partly explain the superiority of this method over the baseline. At inference time, the multilingual sentence encoder approach is however much cheaper than the TRANSLATE TEST baseline, and this method also does not require any machine translation system. Interestingly, the two points difference in accuracy between X-BiLSTM-last and X-BiLSTM-max is maintained across languages, which suggests that having a stronger encoder in English also positively impacts the transfer results on other languages.

	fr	ru	zh
$ft = 1, \lambda = 0.25$ [default]	68.9	66.4	67.9
$ft = 1, \lambda = 0.0$ (no negatives)	67.8	66.2	66.3
$ft = 1, \lambda = 0.5$	64.5	61.3	63.7
$ft = 0, \lambda = 0.25$	68.5	66.3	67.7

Table 5: **Validation accuracy using BiLSTM-max.** Default setting corresponds to $\lambda = 0.25$ (importance of the negative terms) and uses fine-tuning of the target lookup table (ft =1).

For X-BILSTM French, Urdu and Arabic encoders, we plot in Figure 2 the evolution of XNLI dev accuracies and the alignment losses during training. The latter are computed using XNLI parallel dev sentences. We observe a strong correlation between the alignment losses and XNLI accuracies. As the alignment on English-Arabic gets better for example, so does the accuracy on

XNLI-ar. One way to understand this is to recall that the English classifier takes as input the vector $[u, v, |u - v|, u * v]$ where u and v are the embeddings of the premise and hypothesis. So this correlation between $\mathcal{L}_{\text{align}}$ and the accuracy suggests that, as English and Arabic embeddings $[u_{\text{en}}, v_{\text{en}}]$ and $[u_{\text{ar}}, v_{\text{ar}}]$ get closer for parallel sentences (in the sense of the L2-norm), the English classifier gets better at understanding Arabic embeddings $[u_{\text{ar}}, v_{\text{ar}}, |u_{\text{ar}} - v_{\text{ar}}|, u_{\text{ar}} * v_{\text{ar}}]$ and thus the accuracy improves. We observe some over-fitting for Urdu, which can be explained by the small number of parallel sentences (64k) available for that language.

In Table 5, we report the validation accuracy using BiLSTM-max on three languages with different training hyper-parameters. Fine-tuning the embeddings does not significantly impact the results, suggesting that the LSTM alone is ensuring alignment of parallel sentence embeddings. We also observe that the negative term is not critical to the performance of the model, but can lead to slight improvement in Chinese (up to 1.6%).

6 Conclusion

A typical problem in industrial applications is the lack of supervised data for languages other than English, and particularly for low-resource languages. Since annotating data in every language is not a realistic approach, there has been a growing interest in cross-lingual understanding and low-resource transfer in multilingual scenarios. In this work, we extend the development and test sets of the Multi-Genre Natural Language Inference Corpus to 15 languages, including low-resource languages such as Swahili and Urdu. Our dataset, dubbed XNLI, is designed to address the lack of standardized evaluation protocols in cross-lingual understanding, and will hopefully help the community make further strides in this area. We present several approaches based on cross-lingual sentence encoders and machine translation systems. While machine translation baselines obtained the best results in our experiments, these approaches rely on computationally-intensive translation models either at training or at test time. We found that cross-lingual encoder baselines provide an encouraging and efficient alternative, and that further work is required to match the performance of translation based methods.

Acknowledgments

This project has benefited from financial support to Samuel R. Bowman by Google, Tencent Holdings, and Samsung Research.

References

Željko Agić and Natalie Schluter. 2018. Baselines and test data for cross-lingual inference. LREC.

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

Kunchukuttan Anoop, Mehta Pratik, and Bhattacharyya Pushpak. 2018. The iit bombay english-hindi parallel corpus. In LREC.

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In ACL, pages 451–462.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In ICLR.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In EMNLP.

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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Sarath Chandar, Mitesh M. Khapra, Balaraman Ravindran, Vikas Raykar, and Amrita Saha. 2013. Multilingual deep learning. In NIPS, Workshop Track.

Pi-Chuan Chang, Michel Galley, and Christopher D Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In Proceedings of the third workshop on statistical machine translation, pages 224–232.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In ICML, pages 160–167. ACM.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. LREC.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In EMNLP, pages 670–680.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jegou. 2018. Word translation without parallel data. In ICLR.

Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. 2017. An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification. IEEE Journal of Selected Topics in Signal Processing, pages 1340–1348.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using monolingual correlation. In EACL.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. ICLR.

S. Gouews, Y. Bengio, and G. Corrado. 2014. Bilbowa: Fast bilingual distributed representations without word alignments. In ICML.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. LREC.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. NAACL.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In ACL, pages 58–68.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Fine-tuned language models for text classification. arXiv preprint arXiv:1801.06146.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: enabling zero-shot translation. arXiv preprint arXiv:1611.04558.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In Advances in neural information processing systems, pages 3294–3302.

A. Klementiev, I. Titov, and B. Bhattarai. 2012. Inducing crosslingual distributed representations of words. In Coling.

T. Kociský, K.M. Hermann, and P. Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In ACL, pages 224–229.

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

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In ICLR.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In ICML, pages 1188–1196.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2011. Using bilingual parallel corpora for cross-lingual textual entailment. In ACL, HLT '11, pages 1336–1345, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In NIPS, pages 3111–3119.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. J. Artif. Int. Res., 55(1):95–130.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In EMNLP, pages 670–679.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In NAACL, volume 1, pages 2227–2237.

Hieu Pham, Minh-Thang Luong, and Christopher D. Manning. 2015. Learning distributed representations for multilingual text sequences. In Workshop on Vector Space Modeling for NLP.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In Proceedings of *SEM.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. ICLR.

Rafael Sabatini. 1922. Captain Blood. Houghton Mifflin Company.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In LREC, pages 3548–3551.

Holger Schwenk, Ke Tran, Orhan Firat, and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In ACL workshop, Repl4NLP.

Laura Smith, Salvatore Giorgi, Rishi Solanki, Johannes C. Eichstaedt, H. Andrew Schwartz, Muhammad Abdul-Mageed, Anneke Buffone, and Lyle H. Ungar. 2016. Does 'well-being' translate on twitter? In EMNLP, pages 2042–2047.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929–1958.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In NIPS, pages 3104–3112.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In LREC, Istanbul, Turkey. European Language Resources Association (ELRA).

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In LREC.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In IJCAI.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. ICLR.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. In NAACL.

Min Xiao and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In CoNLL, pages 119–129.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. NAACL.

Yuan Zhang, David Gaddy, Regina Barzilay, and Tommi Jaakkola. 2016. Ten pairs to tag–multilingual pos tagging via coarse mapping between embeddings. In NAACL, pages 1307–1317.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In ACL.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In LREC.