Baselines and Test Data for Cross-Lingual Inference

Željko Agić **Natalie Schluter**
Department of Computer Science
IT University of Copenhagen
Rued Langgaards Vej 7, 2300 Copenhagen S, Denmark
zeag@itu.dk nael@itu.dk

Abstract

The recent years have seen a revival of interest in textual entailment, sparked by i) the emergence of powerful deep neural network learners for natural language processing and ii) the timely development of large-scale evaluation datasets such as SNLI. Recast as natural language inference, the problem now amounts to detecting the relation between pairs of statements: they either contradict or entail one another, or they are mutually neutral. Current research in natural language inference is effectively exclusive to English. In this paper, we propose to advance the research in SNLI-style natural language inference toward multilingual evaluation. To that end, we provide test data for four major languages: Arabic, French, Spanish, and Russian. We experiment with a set of baselines. Our systems are based on cross-lingual word embeddings and machine translation. While our best system scores an average accuracy of just over 75%, we focus largely on enabling further research in multilingual inference.

Keywords: natural language inference, cross-lingual methods, test data

1. Introduction

Natural language processing is marking a very recent resurgence of interest in textual entailment. Now revamped as **natural language inference** (NLI) by Bowman et al. (2015) with their SNLI dataset, the task of differentiating contradictory, entailing, and unrelated pairs of sentences (Fig. 1) has entertained a large number of proposals.[1] The timely challenge lends itself to various deep learning approaches such as by Rocktäschel et al. (2015), Parikh et al. (2016), or Wang et al. (2017), which mark a string of very notable results.

Yet, the SNLI corpus is in English only. As of recently, it includes more test data from multiple genres,[2] but it remains exclusive to English. Following Bender (2009) in seeking true language independence, we propose to extend the current NLI research beyond English, and further into the majority realm of low-resource languages.

Since training data is generally unavailable for most languages, work on transfer learning is abundant for the basic NLP tasks such as tagging and syntactic parsing (Das and Petrov, 2011; Ammar et al., 2016). By contrast, the research in cross-lingual entailment is not as plentiful (Negri et al., 2013). To the best of our knowledge, at this point there are no contributions to SNLI-style cross-lingual inference, or for that matter, work on languages other than English at all.

Contributions. In the absence of training data for languages other than English, we propose a set of baselines for **cross-lingual neural inference**. We adapt to the target languages either by i) employing multilingual word embeddings or alternatively by ii) translating the input sentebces into English.

We create **multilingual test data** to facilitate evaluation by manually translating $4 \times 1{,}332$ premise-hypothesis sentence pairs from the English SNLI test data into four other major languages: Arabic, French, Russian, and Spanish.

premise	Female gymnasts warm up before a competition.
entailment	Gymnasts get ready for a competition.
contradiction	Football players practice.
neutral	Gymnasts get ready for the biggest competition of their life.

Figure 1: Example sentence 4-tuple from the SNLI test set, lines 758–760.

We also experiment with automatic translations of the SNLI test data to serve as a proxy for large-scale evaluations in the absence of manually produced data.

2. Cross-Lingual Inference

Following the success of neural networks in SNLI-style inference, we take the neural attention-based model of Parikh et al. (2016) as our starting point. To date, their system remains competitive with the current state of the art. As their attention model is based solely on word embeddings, and is independent of word order, it is particularly suitable for the baseline we present here: a purely multi-lingual embeddings based cross-lingual NLI system. Moreover, their approach is computationally much leaner than most competitors, making it a fast and scalable choice.[3]

In short, the Parikh et al. (2016) model sends sentence pairs, i.e., premises and hypotheses, through a neural pipeline that consists of three separate components:

i) **ATTENTION:** Scores combinations of pairs of words across input sentence pairs. Scores of these word pairs are given by a feed-forward network with ReLU activations that is assumed to model a homomorphic function for linear-time computation. Attention weights for phrases softly aligned with a word are obtained by summing their component vectors each factored by their normalized score.

[1] https://nlp.stanford.edu/projects/snli/

[2] https://repeval2017.github.io/shared/

[3] For more details, see the original paper, and an illustrative overview of the model: https://explosion.ai/blog/deep-learning-formula-nlp.

ii) **COMPARISON:** Word vectors and their aligned phrase counterparts are compared and combined into a single vector using a feed-forward neural network.

iii) **CONCATENATION:** A network that sums over the above output vectors for each input sentence, concatenates this representation and feeds it through a final feed-forward network followed by a linear layer.

To be trained, the model expects SNLI annotations, and an ideally very large vocabulary of distributed word representations.

In this paper, we have at our disposal only a large training corpus of English NLI examples, but a distinct language in which we want to predict for NLI: the target language. We train the system described above on the English training set. We exploit the fact that the system is purely embeddings-based and train with multilingual embeddings for a set of languages including English and the prediction language. Multilingual embeddings are sets of word embeddings generated for multiple languages where the embeddings from the union of these sets are meant to correspond to one another semantically independent of the language the words the embeddings correspond to actually belong. At prediction time, we can safely use the embeddings of the target language.

Mapping. One method for obtaining multilingual word embeddings is to apply the translation matrix technique to a set of monolingual embeddings (Mikolov et al., 2013a) with the aid of a bilingual dictionary containing the source-target word pairs. The method works by finding a transformation matrix from the target language monolingual embeddings to the English monolingual embeddings that minimizes the total least-squared error. This transformation matrix can then be used on words not seen in the bilingual dictionary.

Multilingual embeddings. If parallel sentences or even just parallel documents are available for two or more languages, we can use this data to embed their vocabularies in a shared representation. For example, through an English-Russian parallel corpus we would represent the words of the two languages in a shared space.

There are several competing approaches to training word embeddings over parallel sentences. In this paper, we experiment with four.

BICVM: The seminal approach by Hermann and Blunsom (2014) for inducing bilingual compositional representations from sentence-aligned parallel corpora only.[4]

INVERT: Inverted indexing over parallel corpus sentence IDs as indexing features, with SVD dimensionality reduction on top, following Søgaard et al. (2015) in the recent implementation by Levy et al. (2017).[5] Instead of embedding just language pairs, this method embeds multiple languages into the same space. It is thus distinctly multilingual, rather than just bilingual.

RANDOM: Our implementation of the approach by Vulić and Moens (2016) whereby bilingual SGNS embeddings of Mikolov et al. (2013b) are trained on top of merged pairs of parallel sentences with randomly shuffled tokens.

RATIO: Similar to RANDOM, except the tokens in bilingual sentences are not shuffled, but inserted successively by following the token ratio between the two sentences.

[4] https://github.com/karlmoritz/bicvm

[5] https://bitbucket.org/omerlevy/xling_embeddings/

	ara	**fra**	**spa**	**rus**
eng to ...	25.58	55.80	39.65	30.31
... to **eng**	37.48	46.90	44.04	31.17

Table 1: Machine translation quality (BLEU) for translating the test data from and into English.

Machine translation. One alternative to adapting via shared distributed representations is to use machine translation.

If high-quality translation systems are readily available, or if we can build them from abundant parallel corpora, we can simply translate any input to English and run a pre-trained English NLI model over it. Moreover, we can translate the training data and train target language models similar to Tiedemann et al. (2014) in cross-lingual dependency parsing.

The MT approach only lends itself to medium- to high-density languages. The mapping requires only the monolingual data and bilingual dictionaries, while the bilingual embeddings need parallel texts or documents, both of which are feasible for true low-resource languages.

3. Test Data

The SNLI data are essentially pairs of sentences—premises and hypotheses—each paired with a relation label: contradiction, entailment, or neutral. We had human experts manually translate the first 1,332 test pairs from English into Arabic, French, Russian, and Spanish. We copied over the original labeling of relations, and the annotators manually verified that they hold. That way we can directly evaluate the NLI performance for these five languages.

Further, we translated our test sets into English by Google Translate for our MT-based system as it adapts through translation and thus expects input in English. We also automatically translated the 1,332 original English sentences into our new test languages to check how well we can approximate the "true" accuracies by using translated test data. This way we can facilitate cross-lingual NLI evaluations on a larger scale.

The BLEU scores for the two translation directions are given in Table 1, where we see a clear split by similarity as the translations tend to be better between English, French, and Spanish, and worse outside that group.

4. Experiment

Our experiment involves adapting a neural NLI classifier through multilingual word embeddings and machine translation. We run the Kim et al. (2017) implementation of the attention-based system of Parikh et al. (2016).[6] All models are trained for 15 epochs and otherwise with default settings. While this system typically peaks at over 100 epochs,

[6] https://github.com/harvardnlp/struct-attn

	ara	eng	fra	spa	rus
map to eng					
FASTTEXT	55.75	79.74	51.64	51.94	48.59
bilingual					
BICVM	56.82	76.26	59.03	59.48	54.30
RANDOM	57.35	77.42	63.21	61.01	56.97
RATIO	54.46	78.10	58.64	60.09	51.18
multilingual					
INVERT	54.76	75.10	62.60	60.55	54.76
translation					
FASTTEXT	72.28	–	77.23	75.93	76.54
GLOVE	75.86	–	80.05	78.75	79.59

Table 2: Overall accuracy of the cross-lingual approaches for the target languages and English.

we sacrifice some accuracy to provide more data points in the comparison given the time constraints.

We set the dimensionality to 300 for all our embeddings. Other than that, they are trained with their default settings. In mapping we use the pretrained FASTTEXT vectors[7] for all five languages (Bojanowski et al., 2016). We map the target language embeddings to English as Mikolov et al. (2013a), using the Dinu et al. (2014) implementation[8] and Wiktionary data.[9]

We train our bilingual embeddings on the UN corpus (Ziemski et al., 2016). The corpus covers English and the four target languages with 11M sentences each. The sentences are aligned across all five languages. The Moses tokenizer[10] (Koehn et al., 2007) was used to preprocess the corpus and the test data for training and evaluation.

In the MT approach, we only experiment with translating the input, and not with translating the training data due to time constraints. There, we use two English SNLI models: one with FASTTEXT and the other with GLOVE 840B embeddings (Pennington et al., 2014).[11]

Results. We report the overall accuracy and F_1 scores for the three labels. Table 2 gives the overall scores of our cross-lingual NLI approaches. In general, the more resources we have, the better the scores: Training bilingual embeddings surpasses the mapping to English, while translating to English using a top-level MT system tops the adaptation via embeddings.

The mapping to English works slightly better for Arabic than for the other languages, and scores an average of 52%. The RANDOM bilingual embeddings top their group with an average accuracy of 59.6% followed by INVERT at 58.1%, while RATIO and BICVM are below at 56.1 and 57.4%. The MT approach expectedly tops the table at 75.5% accuracy.

In Table 3 we see that our best bilingual embeddings system RANDOM has a preference for entailment, with ca 9% in F_1 over the other two labels, which makes sense for a model aimed at capturing semantic similarity. This also holds true for the original Parikh et al. (2016) evaluation on English.

	con	ent	neu
ara	55.82	64.17	50.91
fra	57.63	68.73	61.72
spa	55.78	66.98	57.80
rus	56.83	60.61	53.29

Table 3: F_1 scores for **con**tradiction, **ent**ailment, and **neu**tral for our best system, RANDOM.

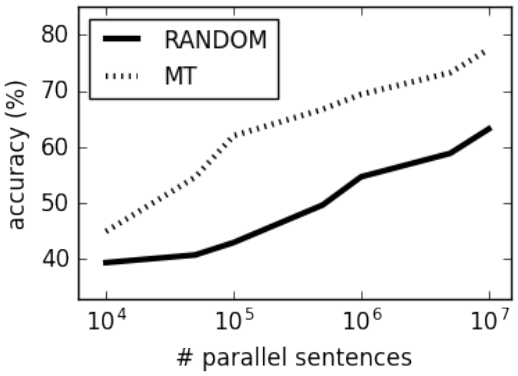

Figure 2: French NLI accuracy in relation to parallel corpus size for RANDOM embeddings.

We report all our English scores as a sanity check. In 100 training epochs, Parikh et al. (2016) score 86.8% with GLOVE 840B as their top score, while we mark 83.4% in 15 epochs. With the significantly smaller FASTTEXT embeddings we reach an accuracy of 79.7%. The multilingual embeddings average at 76.7% for English, where RATIO peaks at 78.1%, likely as its sequential shuffling of parallel texts most closely captures the English sentence structure.

Discussion. Figure 2 plots a learning curve for the French RANDOM approach. We see that its accuracy steadily increases by adding more parallel data into building the bilingual embeddings. As a side note, the MT-based system benefits if the English side of the embeddings grows in size and quality. The figure points out that i) adding more data benefits the task, and that ii) the accuracy of our RANDOM approach stabilizes at around 1M parallel sentences. As per Søgaard et al. (2015) most language pairs can offer no more than 100k sentence pairs, this puts forth a challenge for future cross-lingual NLI learning research.

Replacing the manually prepared test sets with the ones automatically translated from English underestimates the true accuracy by absolute -2.57% on average. The higher the translation quality, the better the estimates we observe: While the difference is around -1% for French and Spanish, it is -7% for Arabic. Still, in proxy evaluation, as with our MT-based adaptation approach in general, we exercise caution: SNLI sentences are image captions, mostly $\leq$15 words long and thus relatively easy to translate (cf. Bowman et al. (2015), Fig. 2) in comparison to, e.g., newspaper text.

[7] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

[8] http://clic.cimec.unitn.it/~georgiana.dinu/down/

[9] https://dumps.wikimedia.org/

[10] https://github.com/moses-smt/mosesdecoder/

[11] https://nlp.stanford.edu/projects/glove/

5. Related Work

Prior to SNLI, there has been work in cross-lingual textual entailment using parallel corpora (Mehdad et al., 2011) and lexical resources (Castillo, 2011), or crowdsourcing for multilingual training data by Negri et al. (2011). We also note two shared tasks, on cross-lingual entailment with five languages (Negri et al., 2013) and English relatedness and inference (Marelli et al., 2014).

Cer et al. (2017) provide multilingual evaluation data within a shared task in semantic textual similarity. There, paired snippets of text are evaluated for their degree of equivalence, and could thus be treated as a fine-grained proxy for SNLI-style evaluations.

SNLI is the first large-scale dataset for NLI in English (Bowman et al., 2015), two orders of magnitude larger than any predecessor. It was recently expanded with test data for multiple genres of English to allow for cross-domain evaluation.[12] Prior to our work, there have been no SNLI-style cross-lingual methods or evaluations.

6. Conclusions

We have proposed the first set of cross-lingual approaches to natural language inference, together with novel test data for four major languages. In experiments with three types of transfer systems, we record viable scores, while at the same time exploring the scalability of cross-lingual inference for low-resource languages.

We are actively enlarging the test data and introducing new languages. Our multilingual test sets and word embeddings are freely available.[13]

7. References

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. (2016). Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Bender, E. M. (2009). Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *EMNLP*, pages 632–642.

Castillo, J. J. (2011). A wordnet-based semantic approach to textual entailment and cross-lingual textual entailment. *International Journal of Machine Learning and Cybernetics*, 2(3):177–189.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval*, pages 1–14.

[12] https://www.nyu.edu/projects/bowman/multinli/

[13] https://bitbucket.org/nlpitu/xnli

Das, D. and Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *ACL*, pages 600–609.

Dinu, G., Lazaridou, A., and Baroni, M. (2014). Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Hermann, K. M. and Blunsom, P. (2014). Multilingual models for compositional distributed semantics. In *ACL*, pages 58–68.

Kim, Y., Denton, C., Hoang, L., and Rush, A. M. (2017). Structured attention networks. *arXiv preprint arXiv:1702.00887*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL*, pages 177–180.

Levy, O., Søgaard, A., and Goldberg, Y. (2017). A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *EACL*, pages 765–774.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., and Zamparelli, R. (2014). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *SemEval*, pages 1–8.

Mehdad, Y., Negri, M., and Federico, M. (2011). Using bilingual parallel corpora for cross-lingual textual entailment. In *ACL*, pages 1336–1345.

Mikolov, T., Le, Q., and Sutskever, I. (2013a). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Negri, M., Bentivogli, L., Mehdad, Y., Giampiccolo, D., and Marchetti, A. (2011). Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *EMNLP*, pages 670–679.

Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., and Giampiccolo, D. (2013). Semeval-2013 task 8: Cross-lingual textual entailment for content synchronization. In *SemEval*, pages 25–33.

Parikh, A., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. In *EMNLP*, pages 2249–2255.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiskỳ, T., and Blunsom, P. (2015). Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Søgaard, A., Agić, v., Martínez Alonso, H., Plank, B., Bohnet, B., and Johannsen, A. (2015). Inverted indexing for cross-lingual nlp. In *ACL*, pages 1713–1722.

Tiedemann, J., Agić, Ž., and Nivre, J. (2014). Treebank

translation for cross-lingual parser induction. In *CoNLL*, pages 130–140.

Vulić, I. and Moens, M.-F. (2016). Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Wang, Z., Hamza, W., and Florian, R. (2017). Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In *LREC*, pages 3530–3534.