Multilingual seq2seq training with similarity loss for cross-lingual document classification

Katherin Yu
Facebook AML
yukatherin@fb.com

Haoran Li
Facebook AML
aimeeli@fb.com

Barlas Oguz
Facebook AML
barlaso@fb.com

Abstract

In this paper we continue the line of work where neural machine translation training is used to produce joint cross-lingual fixed-dimensional sentence embeddings. In this framework we introduce a simple method of adding a loss to the learning objective which penalizes distance between representations of bilingually aligned sentences. We evaluate cross-lingual transfer using two approaches, cross-lingual similarity search on an aligned corpus (Europarl) and cross-lingual document classification on a recently published benchmark Reuters corpus, and we find the similarity loss significantly improves performance on both. Our cross-lingual transfer performance is competitive with state-of-the-art, even while there is potential to further improve by investing in a better in-language baseline. Our results are based on a set of 6 European languages.

1 Introduction

Many real-world services collect data in many languages, and machine learning models on text need to support these languages. In practice, however, it is often only the top one or two dominant languages (usually English) which are supported because it is expensive to collect labeled training data for the task in every language. It is desirable, therefore, to obtain a representation of sequences of text that is joint across all languages, which allows for cross-lingual transfer on the languages without labeled data.

These representations typically take the form of a fixed-size embedding representing a complete sentence or document. Previous work has focused on several approaches in this setting, all of which rely on parallel corpora. In (AP et al., 2013), a predictive auto-encoder is used to reconstruct the featurized representation of a pair of sentences. (Hermann and Blunsom, 2014) constructs a bilingual sentence embedding by minimizing the squared distance between the embeddings of parallel sentences. (Pham et al., 2015) learns a common representation by simultaneously predicting n-grams in both languages from a common vector. In (Mogadala and Rettinger, 2016), a similarity measure is used to minimize distance on both the sentence embeddings, and the average of the word embeddings of a pair of sentences. A method is also proposed to apply this approach to label-aligned corpora in the absence of sentence-aligned corpora by doing a pre-alignment.

Finally, multilingual representations can be learned using a sequence-to-sequence encoder-decoder neural machine translation (NMT) architecture, such as the one introduced in (Sutskever et al., 2014). Multilingual encoders have been successfully demonstrated in the NMT setting (Dong et al., 2015; Firat et al., 2017, 2016; Johnson et al., 2016). Recently (Schwenk et al., 2017) has proposed using this framework for generating multilingual sentence representations and apply it to cross-lingual document classification.

In this paper, we combine this NMT approach with the pairwise similarity approach to obtain better representations. In section 2 we describe our framework. Then in section 4 we present an evaluation of our method based on measuring similarity on the multiply aligned Europarl corpus (Koehn, 2005). Section 5 contains our cross-lingual document classification experiments on the balanced version of the Reuters Corpus Volume 2 dataset (RCV2b), recently published by resampling from the Reuters Corpus Volume 2 to have a balanced distribution of languages and a similar label distribution for each language (Schwenk and Li, 2018).

Proceedings of the 3rd Workshop on Representation Learning for NLP, pages 175–179
Melbourne, Australia, July 20, 2018.

Table 1: SentEval results: performance as a sentence encoder in English

Method	SST	MR	CR	MPQA	SUBJ	TREC	Average
(Conneau et al., 2017) BLSTM, maxpool	81.1	86.3	92.4	90.2	84.6	88.2	87.1
ours, with similarity, meanpool	80.3	73.9	77.5	85.6	90.9	88.0	82.7
ours, with similarity, maxpool	80.4	75.0	79.6	87.3	91.1	88.0	83.56
ours, with similarity, self-attention	80.2	74.3	84.3	88.0	91.8	93.1	85.28

2 Multilingual encoder with similarity loss

We build mostly on the work of (Schwenk et al., 2017) of training an encoder to produce a fixed-dimensional vector representation based on an aggregation over the encoder hidden states. Our setup involves a single shared encoder and decoder with six languages: English, German, French, Spanish, Italian, and Portuguese. We pair languages with English and Spanish, giving 10 unique pairings. The shared vocabulary is of size 85k.

The encoder consists of a two-layer LSTM with hidden sizes 512 and 1024, where the first layer is bidirectional. The decoder is an LSTM without attention, with hidden size 1024. Sentence representations will thus be 1024-dimensional.

We follow the method of prepending a token representing the target language as a first input for the decoder (Johnson et al., 2016). This avoids target-language specific encoder representations since the target language token is not an input to the encoder. We use gradient clipping with max norm 5. We use multi-cca trained word embeddings (Ammar et al., 2016) and allow trainable word embeddings.

2.1 Bilingual batch sampling

Our approach relies on bilingually aligned data. We do not assume multiply aligned (n-way parallel) data, even though we have it in training corpora such as Europarl. Inspired by the m:1 approach in (Schwenk et al., 2017), we train translation in both directions in each batch of bilingually aligned data.

2.2 Translation and similarity loss

We use the average over encoder hidden states to initialize the decoder, and also as a constant input to the decoder at each position, without using attention. The decoder then produces a probability distribution $p_d(t|h)$ on the space of output sequences conditioned on the output of the encoder. Given a set of translation pairs (s, t), let $h(s)$ be the sentence embedding, an elementwise mean of the hidden states of the encoder. The translation loss penalizes the negative log likelihood of the target sequence, given the source:

$$L_{NMT} = \frac{1}{n_t} \sum_{j=1}^{n_t} -\log p_d(t_j|t_1, \cdots, t_{j-1}; h(s))$$

Meanwhile the similarity loss directly minimizes the distance between the embeddings of s and t:

$$L_{sim} = \|h(s) - h(t)\|_2^2$$

We combine these into our final loss term, adding weight regularization on the encoder:

$$L = (L_{NMT}^{src \to tgt} + L_{NMT}^{tgt \to src}) + \alpha L_{sim},$$

where α needs to be chosen to balance the contributions from each term. Note that similarity loss by itself would have a degenerate solution, which is to map all inputs to a constant embedding vector. Introducing negative sampling or a contrastive loss would improve this situation. Note also that both the similarity loss has a regularization effect on the encoder weights. We also try replacing similarity loss term with an L2 norm on the encoder weights. We believe that regularizing encoder weights is important for cross-lingual transfer in that it helps prevent the encoder from "splitting" its output space by source language distribution.

The choice of α depends on relative batch / weight normalization, the distribution of initial word embeddings, hidden size, and other factors. We find that starting with the two terms having comparable value is a good place to start tuning. We tune these parameters to one cross-lingual transfer task (Europarl similarity between De, En, Es).

Training takes about 1.5 days on 4 GPUs for 6 languages with 10 directions. All results are using a single trained encoder in with- and without-similarity loss settings.

Table 2: Europarl (5k) similarity retrieval accuracy from training { without encoder regularization / with encoder weight regularization / with similarity loss }. Some combinations are omitted for space.

	Retrieved language						
	De	En	Es	Fr	It	Pt	All
De	(96.9 / 96.9 / 96.8)	87.0 / 89.2 / 89.8	86.7 // 90.0	85.3 // 89.4	83.2 // 87.2	85.8 // 90.1	87.5 / 89.4 / 90.6
En	85.5 / 89.1 / 88.3	(97.2 / 97.1 / 97.2)	89.9 // 92.4	88.3 // 91.3	86.1 // 89.9	89.4 // 92.0	89.4 / 91.6 / 91.9
Es	85.4 / 87.8 / 87.8	90.2 / 92.0 / **92.4**	(97.1 // 97.0)	88.8 // 91.6	87.5 // **90.9**	91.1 // **93.2**	90.0 / 91.9 / 92.2
Fr	83.8 / 87.4 / 87.8	88.9 / 91.1 / 91.9	89.0 // 92.1	(97.0 // 97.0)	86.2 // 89.8	89.1 // 91.8	89.0 / 91.4 / 91.8
It	82.2 / 85.3 / 85.9	86.7 / 89.4 / 90.3	87.7 // 90.8	86.6 // 90.0	(97.0 // 97.1)	86.9 // 90.9	87.8 / 90.3 / 90.8
Pt	84.5 / 87.6 / **87.9**	90.0 / 91.2 / 92.2	91.0 // **93.0**	88.8 // **91.7**	86.6 // 90.1	(97.3 // 97.3)	89.7 / 91.6 / 92.0
All	86.4 / 89.0 / 89.2	90.0 / 91.7 / 92.3	90.2 // 92.6	89.1 // 91.8	87.8 // 90.8	89.9 // 92.6	88.9 / 91.0 / 91.6

Table 3: Example top 3 retrieved sentences in Europarl 5k: the correctly retrieved sentence is omitted.

Retrieving sentence	*Retrieved (It)*	*Retrieved (Fr)*
Mr President, as it is now Christmas, I would be grateful if you would allow me to speak for a moment.	Signor Presidente, resto in Aula perché mi è stato fatto sapere che, per poter presentare una dichiarazione di voto, occorre essere presenti. Signora Presidente, prendo la parola soltanto per chiedere che, per ragioni ovvie, sia messo a verbale che mi asterrò in questa votazione, visto che mi riguarda in modo diretto.	Monsieur le Président, je reste ici parce que l'on m'a expliqué qu'il fallait être présent dans l'hémicycle pour être autorisé à déposer des explications de vote. Puisque M. Prodi est présent, je vais lui donner la parole en premier, s'il accepte.

3 English performance

We first evaluate our sentence embeddings on a set of English transfer tasks (SentEval). We compare mean pooling, max pooling, and self-attention (Lin et al., 2017) as aggregation methods, with an MLP with one hidden layer of size 128. Our results are several points lower than current best SentEval results.

4 Cross-lingual similarity search

As one of our evaluation methods, we follow (Schwenk et al., 2017) in validating that the closest sentence in an aligned corpus based on our sentence embeddings is the aligned sentence. We use cosine similarity. We use a Europarl development set of 5k sentences across 6 languages and report the accuracy of retrieval in each direction. Note that the corpus has duplicates, thus retrieval cannot be perfect, as reflected in the in-language results. We notice that Portuguese is best for retrieving Spanish sentences and Spanish is best for retrieving Italian and Portuguese sentences.

The results are shown in table 2. As a baseline, we take our setup with NMT loss only, and compare the results with similarity loss added. We see that both encoder weight regularization and similarity loss significantly improve retrieval performance, with similarity loss possibly slightly better.

5 Cross-lingual document classification

One of the main motivations for pursuing multilingual sentence embeddings is to achieve cross-lingual transfer on NLP tasks such as document classification. The multilingual Reuters News Corpus has been adopted as a standard dataset for this task. We will be using a version of this dataset that has been subsampled to obtain even label distribution prior across languages (Schwenk and Li, 2018), to make the interpretation of transfer results easier.

For these tests, we use a linear classifier (logistic regression) and tune the regularization parameter to the development set defined in RCV2Balanced.

5.1 Document segmentation

Method	Mean accuracy
Punctuation, meanpool	74.6
Punctuation, maxpool	67.0
Fixed window, meanpool	73.5
Fixed window, maxpool	68.2

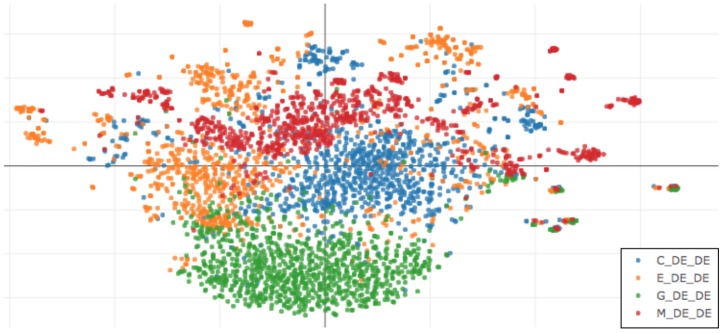
Table 5: Comparison of aggregation methods for document embedding (RCV2Balanced)

Documents in the Reuters corpus are composed of many sentences. In principle, it is possible consider each document as a long sequence and use the resulting embedding from our encoder as-is;

Table 4: Cross-lingual document classification results (RCV2Balanced): from training { without encoder regularization / with similarity loss }. Zero-shot paradigm. Bold indicates best result for target language.

	De	En	Es	Fr	It	Ru	All
De	(91.1 / 90.5)	76.8 / 77.1	67.2 / 76.4	75.3 / **81.7**	63.5 / 71.8	49.5 / 60.5	70.5 / 73.3
En	72.9 / 80.2	(89.0 / 89.4)	72.2 / 74.1	73.0 / 81.0	63.4 / 70.8	60.9 / **65.7**	71.9 / 76.8
Es	76.4 / 79.5	74.1 / 73.4	(92.0 / 92.4)	78.1 / 78.9	68.2 / 72.0	58.7 / 58.0	74.6 / 75.7
Fr	79.5 / **82.5**	79.0 / **80.8**	77.1 / 76.5	(87.5 / 89.9)	68.1 / **72.7**	63.4 / 59.4	75.7 / 77.0
It	76.6 / 78.3	74.8 / 71.2	**76.8** // 75.5	66.8 / 74.0	(81.3 / 81.8)	63.8 / 55.9	73.3 / 72.8
Ru	74.2 / 70.0	72.3 / 71.3	56.3 / 61.8	69.5 / 66.5	64.9 / 60.9	(82.2 / 84.0)	69.9 / 69.1
All	78.4 / 80.2	77.6 / 77.2	73.6 / 76.1	75.0 / 78.7	68.2 / 71.7	63.1 / 63.9	72.7 / 74.6
LOO	- / 78.5	- / (89.4)	- / 73.0	- / 80.5	- / 70.0	- / 65.6	- / (76.2)

Figure 1: t-SNE projection of document embeddings in RCV2Balanced, De test set



however, our encoder would have problems representing such long input sequences with fixed dimensional embeddings, especially because no attention mechanism is present. As a result, we need a method to split a document into smaller sequences, and an aggregation method to go from short sequence embeddings to a document embedding. For splitting we consider simply using the sentences, delimited by punctuation (the characters [.!?]). We also try splitting by a fixed window size (128 words) and fixed stride (64 words). For aggregation, we try elementwise mean- and max-pooling. We find that splitting on punctuation and using mean pooling works best (Table 5).

5.2 Evaluation paradigms

Evaluation paradigm	Mean accuracy
Zero-shot transfer	74.6
Targeted transfer	75.6

Table 6: Comparison of tuning to source- versus target-language development data

Following (Schwenk and Li, 2018), we use two transfer learning paradigms: zero-shot learning and targeted transfer. In zero-shot learning, we tune regularization hyperparameters to the development set in the training/source language and test on the transfer/target language, and the trained model is the same for all directions with the same source; in targeted transfer, we tune these paramaters to the target development set and each model is unique for each dialect direction.

Results are compiled in table 4. It can be seen that adding similarity loss significantly improves over our baseline on average by nearly 2 points. Our best results per target language are better than best results per target language in the zero-shot paradigm in (Schwenk and Li, 2018) using word embeddings and sentence embeddings; however, these are not directly comparable given we are using significantly more training data. Finally, Figure 1 shows a t-SNE representation of the document embeddings over the four classes on a sample of RCVBalanced dataset.

We also try “leaving one out" (LOO) where we pool training data over all languages except the target to augment training data, while tuning to the English development set. However results do not improve over the best single-language transfer numbers (last row in table 4).

6 Conclusion

We presented an improved method for training multi-lingual sentence embeddings, including higher benchmark results for the RCV2 balanced dataset. We showed that including an explicit

similarity loss combined with the encoder-decoder framework improves the quality of multilingual representations. We demonstrated that our representations allow better transfer from one language to another of document classification performance.

We note that although we have shown improvements in RCV2Balanced, our English-only SentEval results are lagging state-of-the-art by at least 2 points. For future work, it is conceivable that starting from a fixed state-of-the-art English encoder (possibly with multitask training with a fixed decoder joint with the English encoder), the similarity loss method could be used to produce the same relative cross-lingual quality while preserving strong in-language performance.

Acknowledgments

We wish to thank Veselin Stoyanov for his mentorship, and our anonymous reviewers for their insightful comments. We look forward to improving this work.

References

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

Sarath Chandar AP, Mitesh M Khapra, Balaraman Ravindran, Vikas Raykar, and Amrita Saha. 2013. Multilingual deep learning. *NIPS DL workshop*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. *Proceedings of ACL-IJNLP*.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T Yarman Vural, and Yoshua Bengio. 2017. Multi-way, multilingual neural machine translation. *Computer Speech & Language*, 45:236–252.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multilingual neural machine translation. *arXiv preprint arXiv:1606.04164*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*.

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*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *eprint arXiv:1703.03130*.

Aditya Mogadala and Achim Rettinger. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–702.

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.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Eleventh International Conference on Language Resources and Evaluation (LREC'18)*. European Language Resources Association (ELRA).

Holger Schwenk, Ke Tran, Orhan Firat, and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *arXiv preprint arXiv: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.