Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond

Presented by: Ife. Adebara

Introduction

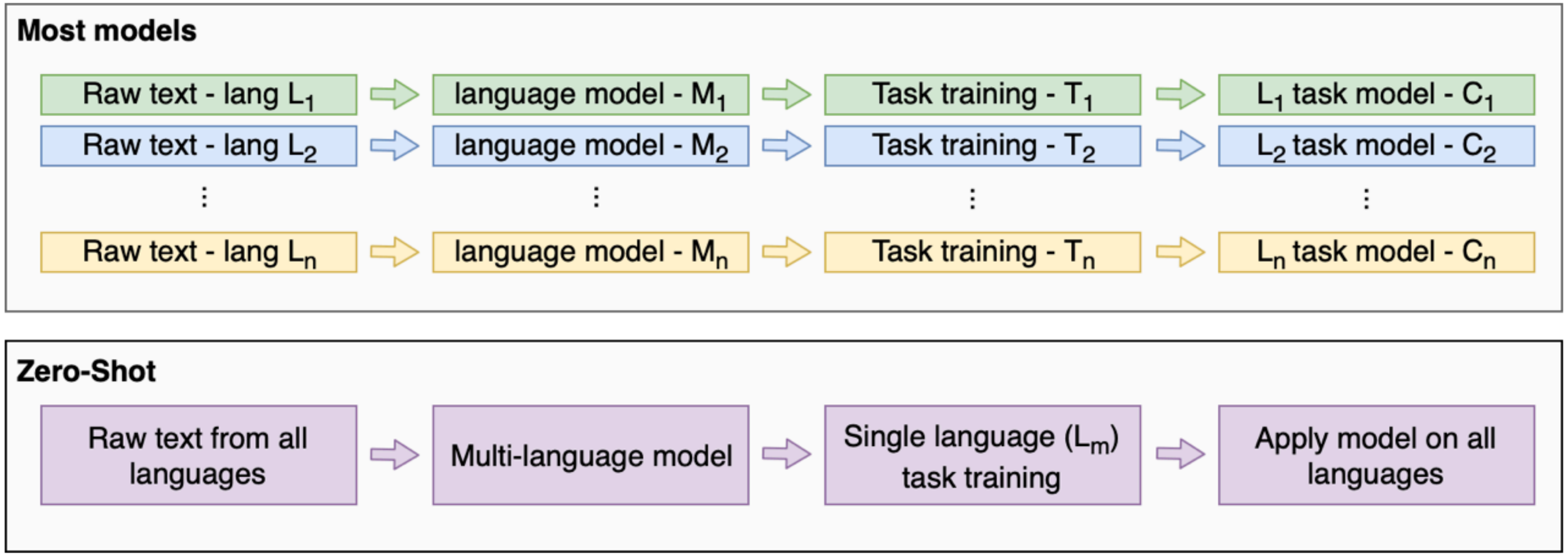
- An architecture to learn joint multilingual sentence representations for 93 languages, belonging to more than 30 different families and written in 28 different scripts.
- Uses a single BiLSTM encoder with a shared BPE vocabulary for all languages. This enables us to learn a classifier on top of the resulting embeddings using English annotated data only, and transfer it to any of the 93 languages without any modification.
- It uses a single encoder to handle multiple languages, so that semantically similar sentences in different languages are close in the embedding space.

Motivation

- NLP techniques are known to be particularly data hungry, limiting their applicability in many practical scenarios.
- Learning a separate model for each language are unable to leverage information across different languages, greatly limiting their potential performance for low-resource languages.
- Languages with limited resources benefit from joint training over many languages, zero-shot transfer of an NLP model from one language to another, and the possibility to handle code-switching.

What is Zero-Shot Transfer

- Zero-Shot learning method aims to solve a task without receiving any example of that task at training phase.
- It can be utilized for a given task by only training the target model (e.g. classifier) on a single language.



Related work

- Learning continuous vector representations of longer linguistic units like sentences (Le and Mikolov, 2014; Kiros et al., 2015)
- Cross-lingual word embeddings (Ruder et al., 2017), which are commonly learned jointly from parallel corpora (Gouws et al., 2015; Luong et al., 2015).
- Separately train word embeddings for each language and map them to a shared space based on a bilingual dictionary (Mikolov et al., 2013a; Artetxe et al., 2018a) or in a fully unsupervised manner (Conneau et al., 2018a; Artetxe et al., 2018b).

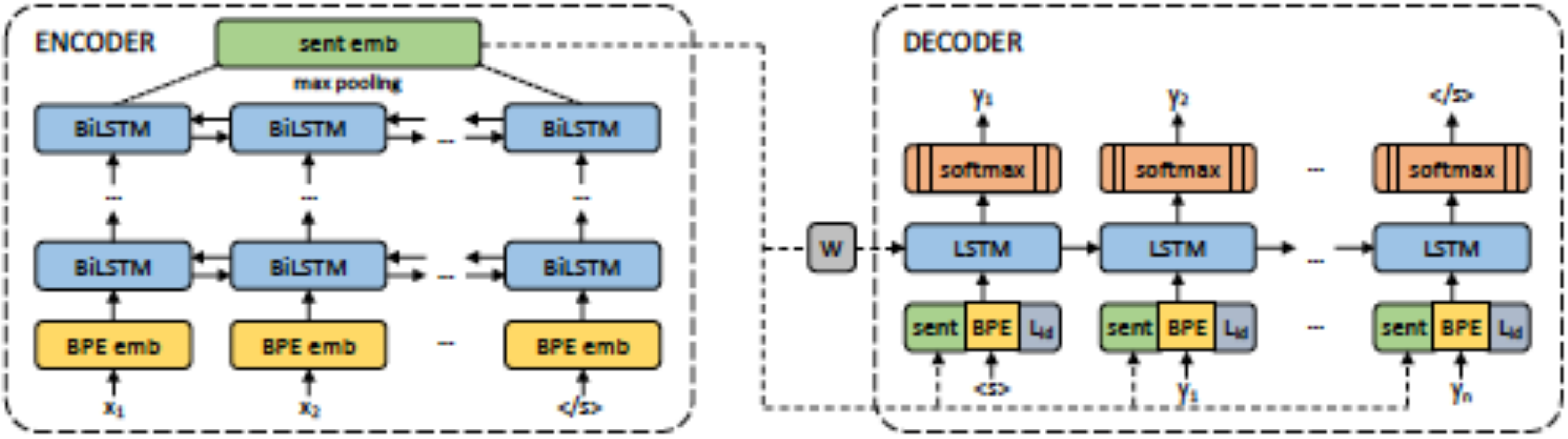


Figure 1: Architecture of our system to learn multilingual sentence embeddings.

Methodology Architecture

- Sequence to sequence encoder decoder
- A single encoder and decoder is shared by all languages involved.
- A joint byte-pair encoding (BPE) vocabulary with 50k operations, which is learned on the concatenation of all training corpora makes the encoder learn language independent representations.
- The decoder takes a language ID embedding that specifies the language to generate, which is concatenated to the input and sentence embed- dings at every time step

Architecture

- A stacked BiLSTM encoder - 1 to 5 layers, each 512-dimensional (after concatenating both directions) are 1024 dimensional.
- Decoder - one layer of dimension 2048.
- The input embedding size is set to 320
- The language ID embedding has 32 dimensions.

Training Strategy

- In preceding work (Schwenk and Douze, 2017; Schwenk, 2018), each input sentence was jointly translated into all other languages.
  - Drawbacks: when trying to scale to a large number of languages, it requires an N-way parallel corpus, which is difficult to obtain for all languages and, it has a quadratic cost with respect to the number of languages
- Therefore, they use only two target languages.
- At the same time, they relax the requirement for N-way parallel corpora by considering separate alignments for each language combination.
- Training minimizes the cross-entropy loss on the training corpus, alternating over all combinations of the languages involved.

Training

- Bi- texts aligned with two target languages – English and Spanish
- Training corpus – combination of Europarl, United Nations, Open-Subtitles2018, Global Voices, Tanzil and Tatoeba corpus (223 million parallel texts in all)
- Preprocessing with Moses; Jieba for Chinese and Mecab for Japanese
- Greek is romanized into the Latin alphabet
- Joint encoder itself has no information on the language or writing script of the tokenized input texts. It is even possible to mix multiple languages in one sentence.

	af	am	ar	ay	az	be	ber	bg	bn	br	bs	ca	cbk	cs	da	de
train sent.	67k	88k	8.2M	14k	254k	5k	62k	4.9M	913k	29k	4.2M	813k	1k	5.5M	7.9M	8.7M
en→xx err.	11.20	60.71	8.30	n/a	44.10	31.20	29.80	4.50	10.80	83.50	3.95	4.00	24.20	3.10	3.90	0.90
xx→en err.	9.90	55.36	7.80	n/a	23.90	36.50	33.70	5.40	10.00	84.90	3.11	4.20	21.70	3.80	4.00	1.00
test sent.	1000	168	1000	–	1000	1000	1000	1000	1000	1000	354	1000	1000	1000	1000	1000

	dtp	dv	el	en	eo	es	et	eu	fi	fr	ga	gl	ha	he	hi	hr
train sent.	1k	90k	6.5M	2.6M	397k	4.8M	5.3M	1.2M	7.9M	8.8M	732	349k	127k	4.1M	288k	4.0M
en→xx err.	92.10	n/a	5.30	n/a	2.70	1.90	3.20	5.70	3.70	4.40	93.80	4.60	n/a	8.10	5.80	2.80
xx→en err.	93.50	n/a	4.80	n/a	2.80	2.10	3.40	5.00	3.70	4.30	95.80	4.40	n/a	7.60	4.80	2.70
test sent.	1000	–	1000	–	1000	1000	1000	1000	1000	1000	1000	1000	–	1000	1000	1000

	hu	hy	ia	id	ie	io	is	it	ja	ka	kab	kk	km	ko	ku	kw
train sent.	5.3M	6k	9k	4.3M	3k	3k	2.0M	8.3M	3.2M	296k	15k	4k	625	1.4M	50k	2k
en→xx err.	3.90	59.97	5.40	5.20	14.70	17.40	4.40	4.60	3.90	60.32	39.10	80.17	77.01	10.60	80.24	91.90
xx→en err.	4.00	67.79	4.10	5.80	12.80	15.20	4.40	4.80	5.40	67.83	44.70	82.61	81.72	11.50	85.37	93.20
test sent.	1000	742	1000	1000	1000	1000	1000	1000	1000	746	1000	575	722	1000	410	1000

	kzj	la	lfn	lt	lv	mg	mhr	mk	ml	mr	ms	my	nb	nds	nl	oc
train sent.	560	19k	2k	3.2M	2.0M	355k	1k	4.2M	373k	31k	2.9M	2k	4.1M	12k	8.4M	3k
en→xx err.	91.60	41.60	35.90	4.10	4.50	n/a	87.70	5.20	3.35	9.00	3.40	n/a	1.30	18.60	3.10	39.20
xx→en err.	94.10	41.50	35.10	3.40	4.70	n/a	91.50	5.40	2.91	8.00	3.80	n/a	1.10	15.60	4.30	38.40
test sent.	1000	1000	1000	1000	1000	–	1000	1000	687	1000	1000	–	1000	1000	1000	1000

	pl	ps	pt	ro	ru	sd	si	sk	sl	so	sq	sr	sv	sw	ta	te
train sent.	5.5M	4.9M	8.3M	4.9M	9.3M	91k	796k	5.2M	5.2M	85k	3.2M	4.0M	7.8M	173k	42k	33k
en→xx err.	2.00	7.20	4.70	2.50	4.90	n/a	n/a	3.10	4.50	n/a	1.80	4.30	3.60	45.64	31.60	18.38
xx→en err.	2.40	6.00	4.90	2.70	5.90	n/a	n/a	3.70	3.77	n/a	2.30	5.00	3.20	39.23	29.64	22.22
test sent.	1000	1000	1000	1000	1000	–	–	1000	823	–	1000	1000	1000	390	307	234

	tg	th	tl	tr	tt	ug	uk	ur	uz	vi	wuu	yue	zh
train sent.	124k	4.1M	36k	5.7M	119k	88k	1.4M	746k	118k	4.0M	2k	4k	8.3M
en→xx err.	n/a	4.93	47.40	2.30	72.00	59.90	5.80	20.00	82.24	3.40	25.80	37.00	4.10
xx→en err.	n/a	4.20	51.50	2.60	65.70	49.60	5.10	16.20	80.37	3.00	25.20	38.90	5.00
test sent.	–	548	1000	1000	1000	1000	1000	1000	428	1000	1000	1000	1000

Table 1: List of the 93 languages along with their training size, the resulting similarity error rate on Tatoeba, and the number of sentences in it. Dashes denote language pairs excluded for containing less than 100 test sentences.

Evaluation

- The model is trained only on sentences in English and tested on all languages. The encoder is also constant and not fine-tuned for every task
  - XNLI - transfer performance of an NLI model trained on English data over 14 additional test languages
  - Cross-lingual document classification (MLDoc)
  - Bitext mining (BUCC
  - Multilingual similarity search in 112 languages

XNLI

- Given two sentences, a premise and a hypothesis, decides whether there is an *entailment*, *contradiction* or *neutral* relationship
- 2,500 development and 5,000 test instances translated from English into 14 languages
- Train a classifier on top of our multilingual encoder using the combination of the two sentence embeddings: $(p, h, p \cdot h, |p - h|)$, where p and h are the premise and hypothesis.
- All hyperparameters were optimized on the English XNLI development corpus only, and then, the same classifier was applied to all languages of the XNLI test set. As such, we did not use any training or development data in any of the foreign languages.
- the multilingual sentence embeddings are fixed and not fine-tuned on the task or the language.

XNLI Evaluation

		EN	EN → XX													
			fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
Zero-Shot Transfer, one NLI system for all languages:																
Conneau et al.	X-BiLSTM	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
(2018b)	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
BERT uncased*	Transformer	81.4	–	74.3	70.5	–	–	–	–	62.1	–	–	63.8	–	–	58.3
Proposed method	BiLSTM	73.9	**71.9**	72.9	72.6	**72.8**	**74.2**	**72.1**	**69.7**	**71.4**	**72.0**	**69.2**	71.4	**65.5**	**62.2**	61.0
Translate test, one English NLI system:																
Conneau et al. (2018b)	BiLSTM	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
BERT uncased*	Transformer	81.4	–	74.9	74.4	–	–	–	–	70.4	–	–	70.1	–	–	**62.1**
Translate train, separate NLI systems for each language:																
Conneau et al. (2018b)	BiLSTM	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
BERT cased*	Transformer	**81.9**	–	**77.8**	**75.9**	–	–	–	–	70.7	–	*68.9*[†]	**76.6**	–	–	61.6

Table 2: Test accuracies on the XNLI cross-lingual natural language inference dataset. All results from Conneau et al. (2018b) correspond to max-pooling, which outperforms the last-state variant in all cases. Results involving MT do not use a multilingual model and are not directly comparable with zero-shot transfer. Overall best results are in bold, the best ones in each group are underlined.

MLDoc: cross-lingual classification

- MLDoc dataset Schwenk and Li (2018),
- 1,000 training and development; 4,000 test documents for each language, divided in 4 different genres
- Train a classifier on top of our multilingual encoder using the English training data, optimizing hyperparameters on the English development set, and evaluating the resulting system in the remaining languages.

		EN	EN → XX						
			de	es	fr	it	ja	ru	zh
Schwenk and Li (2018)	MultiCCA + CNN	**92.20**	81.20	72.50	72.38	69.38	**67.63**	60.80	**74.73**
	BiLSTM (Europarl)	88.40	71.83	66.65	72.83	60.73	-	-	-
	BiLSTM (UN)	88.83	-	69.50	74.52	-	-	61.42	71.97
Proposed method		89.93	**84.78**	**77.33**	**77.95**	**69.43**	60.30	**67.78**	71.93

Table 3: Accuracies on the MLDoc zero-shot cross-lingual document classification task (test set).

BUCC: bitext mining

- Given two comparable corpora in different languages, the task consists in identifying sentence pairs that are translations of each other.
- Extracting parallel sentences from a comparable corpus between English and four foreign languages: German, French, Russian and Chinese.
- The dataset consists of 150K to 1.2M sentences for each language, split into a sample, training and test set, with about 2–3% of the sentences being parallel.

$$\text{score}(x, y) = \text{margin}(\cos(x, y), \sum_{z \in \text{NN}_k(x)} \frac{\cos(x, z)}{2k} + \sum_{z \in \text{NN}_k(y)} \frac{\cos(y, z)}{2k})$$

where x and y are the source and target sentences, and $\text{NN}_k(x)$ denotes the k nearest neighbors of x in the other language. The paper explores different margin functions, with *ratio* ($\text{margin}(a, b) = \frac{a}{b}$) yielding the best results. This notion of margin is related to CSLS (Conneau et al., 2018a).

	TRAIN				TEST			
	de-en	fr-en	ru-en	zh-en	de-en	fr-en	ru-en	zh-en
Azpeitia et al. (2017)	83.33	78.83	-	-	83.74	79.46	-	-
Grégoire and Langlais (2017)	-	20.67	-	-	-	20	-	-
Zhang and Zweigenbaum (2017)	-	-	-	43.48	-	-	-	45.13
Azpeitia et al. (2018)	84.27	80.63	80.89	76.45	85.52	81.47	81.30	77.45
Bouamor and Sajjad (2018)	-	75.2	-	-	-	76.0	-	-
Chongman Leong and Chao (2018)	-	-	-	58.54	-	-	-	56
Schwenk (2018)	76.1	74.9	73.3	71.6	76.9	75.8	73.8	71.6
Artetxe and Schwenk (2018)	94.84	91.85	90.92	91.04	95.58	92.89	92.03	**92.57**
Proposed method	**95.43**	**92.40**	**92.29**	**91.20**	**96.19**	**93.91**	**93.30**	92.27

Table 4: F1 scores on the BUCC mining task.

Tatoeba: similarity search

- The dataset consists of up to 1,000 English-aligned sentence pairs for each language.
- Evaluation is done by finding the nearest neighbor for each sentence in the other language according to cosine similarity and computing the error rate.

Ablation Studies

Depth	Tatoeba Err [%]	BUCC F1	MLDoc Acc [%]	XNLI-en Acc [%]	XNLI-xx Acc [%]
1	37.96	89.95	69.42	70.94	64.54
3	28.95	92.28	71.64	72.83	68.43
5	**26.31**	**92.83**	**72.79**	**73.67**	**69.92**

Table 5: Impact of the depth of the BiLSTM encoder.

NLI obj.	Tatoeba Err [%]	BUCC F1	MLDoc Acc [%]	XNLI-en Acc [%]	XNLI-xx Acc [%]
-	**26.31**	92.83	72.79	73.67	**69.92**
×1	26.89	93.01	**74.51**	73.71	69.10
×2	28.52	**93.06**	71.90	74.65	67.75
×3	27.83	92.98	73.11	**75.23**	61.86

Table 6: Multitask training with an NLI objective and different weightings.

#langs	WMT Err [%]	BUCC F1	MLDoc Acc [%]	XNLI-en Acc [%]	XNLI-xx Acc [%]
All (93)	**0.54**	92.83	72.79	**73.67**	**69.92**
Eval (18)	0.59	**92.91**	**75.63**	72.99	68.84

Table 7: Comparison between training on 93 languages and training on the 18 evaluation languages only.