Language Transfer Learning

13 października 2019

Outline

Byte Pair Encoding

Neural Machine Translation of Rare Words with Subword Units
(Sennrich et al. 2015)

Byte Pair Encoding

In the paper "Neural Machine Translation of Rare Words with Subword Units" published in 2015, the author has released the source code of doing byte pair encoding for a corpus of words. We count the frequency of each word shown in the corpus. For each word, we append a special stop token "</w>" at the end of the word. We will talk about the motivation behind this later. We then split the word into characters. Initially, the tokens of word are all of its characters plus the additional "</w>" token. For example, the tokens for word "low" are ["l", "o", "w", "</w>"] in order. So after counting all the words in the dataset, we will get a vocabulary for the tokenized word with its corresponding counts, such as

```
{'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w e s t </w>': 6, 'w i d e s t </w>': 3}
```

In each iteration, we count the frequency of each consecutive byte pair, find out the most frequent one, and merge the two byte pair tokens to one token.

For the above example, in the first iteration of merge, because byte pair "e" and "s" occurred 6 + 3 = 9 times which is the most frequent. We merge these to into a new token "es". Note that because token "s" is also gone in this particular example.

```
{'l o w </w>': 5, 'l o w e r </w>': 2, 'n e w es t </w>': 6, 'w i d es t </w>': 3}
```

https://leimao.github.io/blog/Byte-Pair-Encoding/

Outline

Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond (2018, Artetxe et al.)

Motywacja

1. wszystkie obecne modele są ”data hungry”
2. warto zrobić transfer learning z angielskiego do innych języków
3. pierwsza praca, która pracuje na 93 językach (z low-resource jezykami)

Sposoby ewaluacji

Nie ma ugruntowionych zbiorów do tego zadania. Ewaluacja na:

- cross-lingual natural language inference XNLI Dataset (15 jezykow + ang)
- cross-lingual classification MLDoc Dataset
- bitext Mining (BUCC dataset)
- nowy task - multilingual similarity search na Tatoeba corpus

Architektura

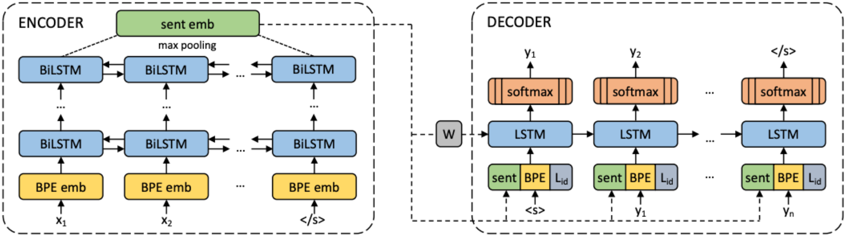

Figure 1: Architecture of our system to learn multilingual sentence embeddings.

wspolny BPE 50k, enkoder nie ma informacji o jezyku
ENKODER stacked-bilstm 1-5 layers, 512 dim (1024 ostatecznie)
DEKODER lstm 1 layer, 2048 dim, jezyk: embedding 32 dim

Trenowanie

- wczesniej korpus rownolegly jezyk-jezyk (problem zlozonosci kwadratowej)
- teraz tylko 2 jezyki ze wszystkimi tłumaczeniami (starczy nawet 1)- angielski i hiszpański
- bez rownoleglych (autoencoding) daje kiepskie wyniki
- korpusy: Europarl, United Nations, OpenSubtitles2018, Global Voices, Tanzil and Tatoeba (93 języki)

XNLI

- NLI - 2 zdania i wybrać (entailment, contradiction, neutral)
- 2500 dev zdan (wszystkie przetlumaczone)
- 5000 test zdan
- trenowanie tylko na angielskim!
- wejscie- $(p, h, ph, |p - h|)$
- wytrenowany liniowy klasyfikator na enkoderze

wyniki na XNLI

		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. (2018b)	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
	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.

* Results for BERT (Devlin et al., 2019) are extracted from its GitHub README[9]

† Monolingual BERT model for Thai from `https://github.com/ThAIKeras/bert`

MLDoc

- 1000 train i dev dokumentow
- 4000 test doc
- 4 kategorie
- tak samo- trenowane tylko na angielskim
- klasyfikator ff, jedna ukryta 10 units

MLDoc-wyniki

		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

- dwa korpusy w roznych jezykach
- nalezy znalezc zdania ktore sa tlumaczeniami

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

- x- source language, y- target language
- $NN_k(x)$- k- najblizszych sasiadow x w drugim jezyku
- rozne ratio $\text{margin}(a, b) = \frac{a}{b}$

BUCC- wyniki

	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

- autorzy wprowadzili
- 112 jezykow
- do 1000 par zdan na kazdy jezyk- ang
- ewaluacja - szukanie najblizszego sąsiada w drugim języku

Tatoeba- wyniki

	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

the number of sentences in it. Dashes denote language pairs excluded for containing less than 100 test sentences.

Outline

Cross-lingual Language Model Pretraining

Cross-lingual Language Model Pretraining (2019, Lample et al.)

shared BPE

- rozkład wielomianowy
- n_i- i-ty język
- $\alpha = 0.5$

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \quad \textbf{with} \quad p_i = \frac{n_i}{\sum_{k=1}^{N} n_k}.$$

Zwiększa nakład na low-resource języki

zadania przy trenowaniu

- Causal Language Modeling (CLM)- standardowo przekazuje się poprzedni hidden state, ale tutaj tego nie robią
- Masked Language Modeling (MLM)- losowej dlugości text stream zamiast par jak w orginalnym BERTcie
- Translation Language Modeling (TLM) - korpus bilingualny

MLM, TLM

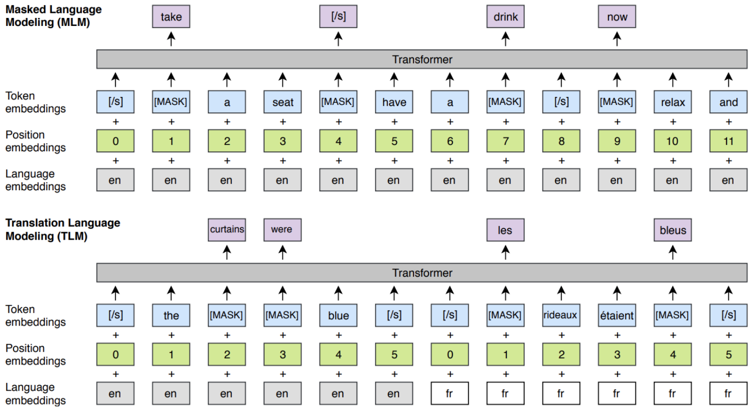

Figure 1: **Cross-lingual language model pretraining.** The MLM objective is similar to the one of Devlin et al. (2018), but with continuous streams of text as opposed to sentence pairs. The TLM objective extends MLM to pairs of parallel sentences. To predict a masked English word, the model can attend to both the English sentence and its French translation, and is encouraged to align English and French representations. Position embeddings of the target sentence are reset to facilitate the alignment.

XNLI

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. (2018b)	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	65.6
Devlin et al. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**

Table 1: **Results on cross-lingual classification accuracy.** Test accuracy on the 15 XNLI languages. We report results for machine translation baselines and zero-shot classification approaches based on cross-lingual sentence encoders. XLM (MLM) corresponds to our unsupervised approach trained only on monolingual corpora, and XLM (MLM+TLM) corresponds to our supervised method that leverages both monolingual and parallel data through the TLM objective. Δ corresponds to the average accuracy.

unsupervised machine translation

		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lample et al. (2018b)							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

Table 2: **Results on unsupervised MT.** BLEU scores on WMT'14 English-French, WMT'16 German-English and WMT'16 Romanian-English. For our results, the first two columns indicate the model used to pretrain the encoder and the decoder. " - " means the model was randomly initialized. EMB corresponds to pretraining the lookup table with cross-lingual embeddings, CLM and MLM correspond to pretraining with models trained on the CLM or MLM objectives.

supervised machine translation

Pretraining	-	CLM	MLM
Sennrich et al. (2016)	33.9	-	-
ro $\rightarrow$ en	28.4	31.5	35.3
ro $\leftrightarrow$ en	28.5	31.5	35.6
ro $\leftrightarrow$ en + BT	34.4	37.0	**38.5**

Table 3: **Results on supervised MT.** BLEU scores on WMT'16 Romanian-English. The previous state-of-the-art of Sennrich et al. (2016) uses both back-translation and an ensemble model. ro $\leftrightarrow$ en corresponds to models trained on both directions.

language modelling

Training languages	Nepali perplexity
Nepali	157.2
Nepali + English	140.1
Nepali + Hindi	115.6
Nepali + English + Hindi	**109.3**

Table 4: **Results on language modeling.** Nepali perplexity when using additional data from a similar language (Hindi) or a distant one (English).