Multilingual Models for Compositional Distributed Semantics

Karl Moritz and Phil Blunsom
University of Oxford

PRESENTED BY

NITHIN HOLLA

MSC AI, UNIVERSITY OF AMSTERDAM

Motivation

- Over 7000 known languages
- Most of NLP focused on English
- Many languages are low-resource
- Universal language learner – a holy grail of NLP?

https://www.ethnologue.com/guides/how-many-languages

Overview

- Learn word embeddings across languages in a shared multilingual semantic space
- Similar words lie close by and dissimilar words are far apart
- Unsupervised approach
- Make use of sentence-aligned parallel corpora

The weather is nice today
Het is lekker weer vandaag

The book was interesting
Het boek was interessant

- Composition Vector Model (CVM) for sentences and documents
- No need of syntactic parse trees, word alignment or annotations
- Downstream tasks can be made language-agnostic

http://www.marekrei.com/blog/multilingual-semantic-models/

Approach

- Parallel sentences share semantics, hence should also share the representation
- Consider two languages x and y and sentence embedding functions $f: X \rightarrow \mathbb{R}^d$ and $g: Y \rightarrow \mathbb{R}^d$
- Let C be the parallel corpus. For two sentences $(a, b) \in C$, the energy is defined as

$$\mathrm{E}_{bi}(a, b) = \|f(a) - g(b)\|^2$$

- For every (a, b) sample sentences n that are not related to a for hinge loss

$$E_{hl}(a, b, n) = \max\Big(\big(m + E_{bi}(a, b) - E_{bi}(a, n)\big), 0\Big)$$

- Final objective function

$$J(\theta) = \sum_{(a,b)\in C} \sum_{i=1}^{k} E_{hl}(a, b, n_i) + \frac{\lambda}{2}\|\theta\|^2$$

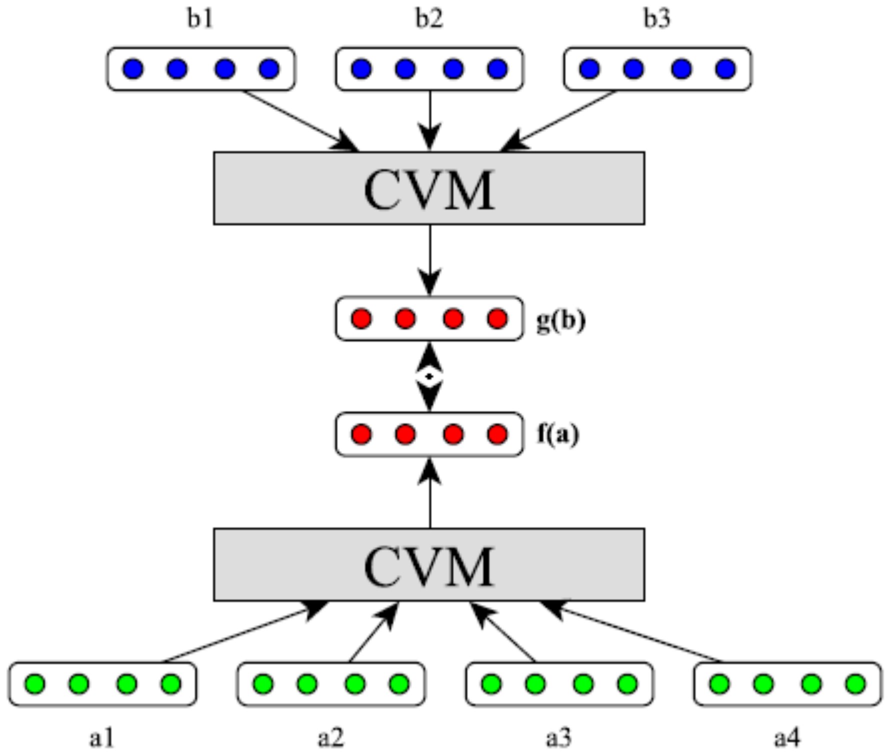
Two Composition Models

- ***ADD*** – Represents a sentence by sum of its word vectors

$$f_{ADD}(x) = \sum_{i=1}^{n} x_i$$

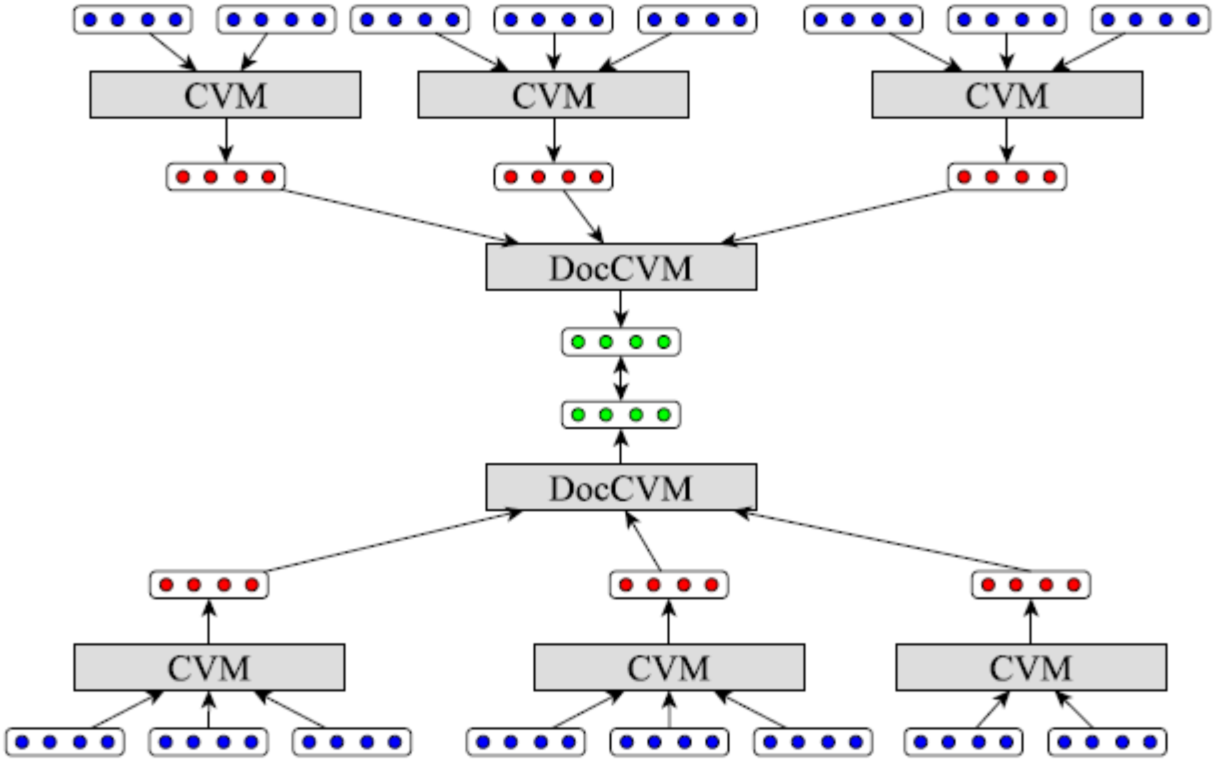
- ***BI*** – Captures interaction with non-linearity over bigram pairs

$$f_{BI}(x) = \sum_{i=1}^{n} \tanh(x_{i-1} + x_i)$$



Document Representation

- Compose sentences into documents
- Recursively apply composition and a similar objective function
- 4 models – ***ADD, BI, DOC/ADD, DOC/BI***



Corpora

Europarl v7

- Parallel corpus extracted from the proceedings of the European Parliament
- 21 European languages
- EN → L2 and L2 → EN

TED Corpus

- English transcripts and translations of TED talks
- Selected subset of talks based on keywords - technology, culture, science, global issues etc.
- Keywords used as document labels for classification task later
- 12,078 parallel documents across 12 language pairs

Training

- Model weights initialized according to Gaussian distribution with $\mu = 0$, $\sigma^2 = 0.1$
- Development set used to set hyperparameters
- For each positive sample, $k \in \{1, 10, 50\}$ noise samples used
- Embedding dimensionality $d = 128$
- Margin $m = d$
- L2 regularization with $\lambda = 0.1$
- Learning rate in $\{0.01, 0.05\}$
- Batch size $b \in \{10, 50\}$
- AdaGrad optimizer

RCV1/RCV2 Document Classification Experiment

- Contains news articles with topic as labels (not parallel corpora)
- Experiment with EN $\rightarrow$ DE and DE $\rightarrow$ EN
- Embeddings first learned from Europarl corpus
- Document represented by average embedding of all its sentences
- Train multiclass classifier using average perceptron
- Training on English and testing on German documents and vice-versa

RCV1/RCV2 Document Classification Experiment

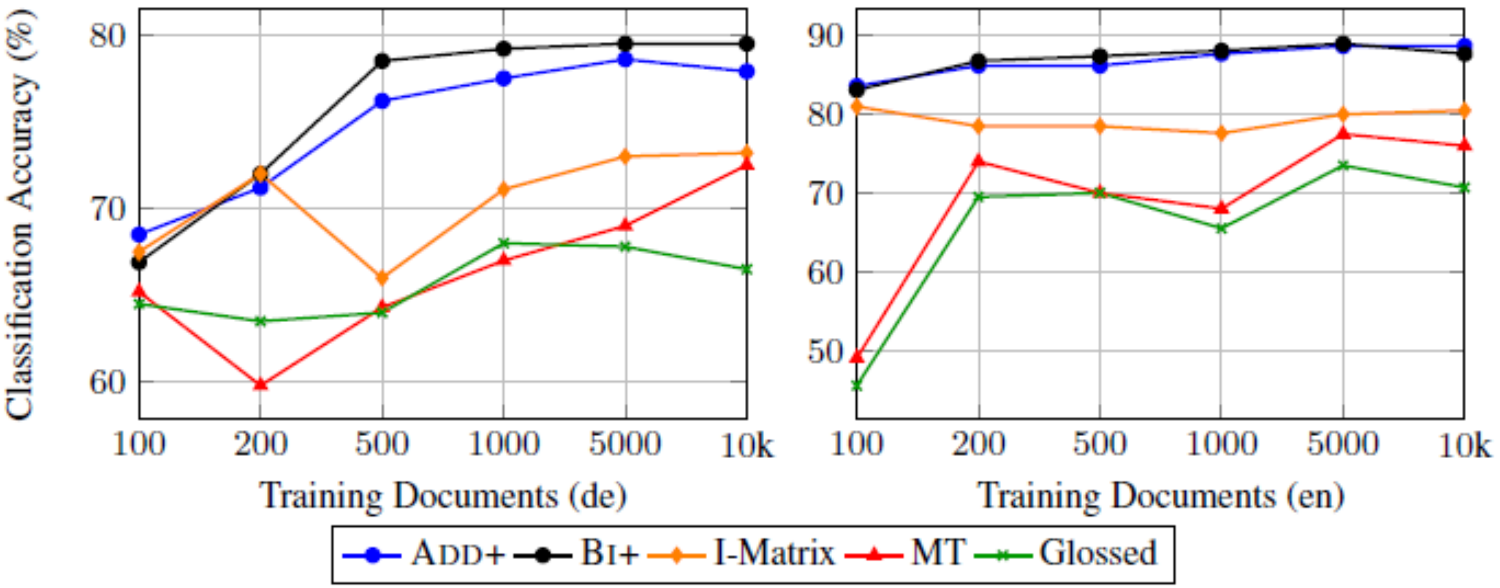
Model	en → de	de → en
Majority Class	46.8	46.8
Glossed	65.1	68.6
MT	68.1	67.4
I-Matrix	77.6	71.1
$dim = 40$		
ADD	83.7	71.4
ADD+	86.2	76.9
BI	83.4	69.2
BI+	86.9	74.3
$dim = 128$		
ADD	86.4	74.7
ADD+	87.7	77.5
BI	86.1	79.0
BI+	**88.1**	**79.2**

Classification accuracy for training on 1000 examples

- ***ADD*** – Trained on 500k sentence pairs
- ***ADD+*** – Trained with addition of 500k EN-FR pairs
- ***BI, BI+*** – Likewise but with bigrams
- ***BI*** models outperform ***ADD*** in general
- French acts like a pivot language and improves performance
- ***BI*** models not always better than ***ADD+***
- EN → DE better than DE → EN

RCV1/RCV2 Document Classification Experiment

- Small performance gain with increasing number of documents
- Decent performance with just 100 documents
- Possibly high bias due to simplistic model



TED Corpus Experiments

- Training performed in two settings:
  - Single mode – Vectors learned from single language pair EN-X
  - Joint mode – Vectors learned from all parallel sub-corpora simultaneously
- ***DOC*** models trained with ***ADD*** and ***BI*** as CVM in single and joint mode
- Document representations used to train classifiers
- Classifier trained on one language and evaluated on another
- Machine translation baseline
  - For the experiment $L_1 \rightarrow L_2$, train Naive Bayes classifier on L_1 and evaluate on translated L_2
  - Expected to be a strong baseline

TED Corpus Experiments

Setting	Arabic	German	Spanish	French	Italian	Dutch	Polish	Pt-Br	Roman.	Russian	Turkish
en → L2											
MT System	**0.429**	**0.465**	**0.518**	**0.526**	**0.514**	**0.505**	**0.445**	**0.470**	**0.493**	0.432	0.409
ADD *single*	0.328	0.343	0.401	0.275	0.282	0.317	0.141	0.227	0.282	0.338	0.241
BI *single*	0.375	0.360	0.379	0.431	0.465	0.421	0.435	0.329	0.426	0.423	**0.481**
DOC/ADD *single*	0.410	0.424	0.383	0.476	0.485	0.264	0.402	0.354	0.418	0.448	0.452
DOC/BI *single*	0.389	0.428	0.416	0.445	0.473	0.219	0.403	0.400	0.467	0.421	0.457
DOC/ADD *joint*	0.392	0.405	0.443	0.447	0.475	0.453	0.394	0.409	0.446	**0.476**	0.417
DOC/BI *joint*	0.372	0.369	0.451	0.429	0.404	0.433	0.417	0.399	0.453	0.439	0.418
L2 → *en*											
MT System	0.448	0.469	**0.486**	0.358	**0.481**	0.463	**0.460**	0.374	**0.486**	0.404	0.441
ADD *single*	0.380	0.337	0.446	0.293	0.357	0.295	0.327	0.235	0.293	0.355	0.375
BI *single*	0.354	0.411	0.344	0.426	0.439	0.428	0.443	0.357	0.426	0.442	0.403
DOC/ADD *single*	**0.452**	**0.476**	0.422	0.464	0.461	0.251	0.400	0.338	0.407	**0.471**	0.435
DOC/BI *single*	0.406	0.442	0.365	**0.479**	0.460	0.235	0.393	0.380	0.426	0.467	**0.477**
DOC/ADD *joint*	0.396	0.388	0.399	0.415	0.461	**0.478**	0.352	**0.399**	0.412	0.343	0.343
DOC/BI *joint*	0.343	0.375	0.369	0.419	0.398	0.438	0.353	0.391	0.430	0.375	0.388

F1-scores for TED document classification task

- MT baseline is often the best but other models not far behind
- ***DOC*** model usually performs better
- Joint model is not always the top-performer
- ***ADD*** models outperform ***BI*** models in many cases

TED Corpus Experiments

Training Language	Test Language										
	Arabic	German	Spanish	French	Italian	Dutch	Polish	Pt-Br	Rom'n	Russian	Turkish
Arabic		0.378	0.436	0.432	0.444	0.438	0.389	0.425	0.420	0.446	0.397
German	0.368		0.474	0.460	0.464	0.440	0.375	0.417	0.447	0.458	0.443
Spanish	0.353	0.355		0.420	0.439	0.435	0.415	0.390	0.424	0.427	0.382
French	0.383	0.366	0.487		0.474	0.429	0.403	0.418	0.458	0.415	0.398
Italian	0.398	0.405	0.461	0.466		0.393	0.339	0.347	0.376	0.382	0.352
Dutch	0.377	0.354	0.463	0.464	0.460		0.405	0.386	0.415	0.407	0.395
Polish	0.359	0.386	0.449	0.444	0.430	0.441		0.401	0.434	0.398	0.408
Portuguese	0.391	0.392	0.476	0.447	0.486	0.458	0.403		0.457	0.431	0.431
Romanian	0.416	0.320	0.473	0.476	0.460	0.434	0.416	0.433		0.444	0.402
Russian	0.372	0.352	0.492	0.427	0.438	0.452	0.430	0.419	0.441		0.447
Turkish	0.376	0.352	0.479	0.433	0.427	0.423	0.439	0.367	0.434	0.411	

F1-scores for TED document classification task

- ***DOC/ADD*** joint model from previous experiment
- Classifier trained and tested on languages without parallel data
- Non-English languages
- Scores similar to previous table indicate embeddings for all languages are useful

TED Corpus Experiments

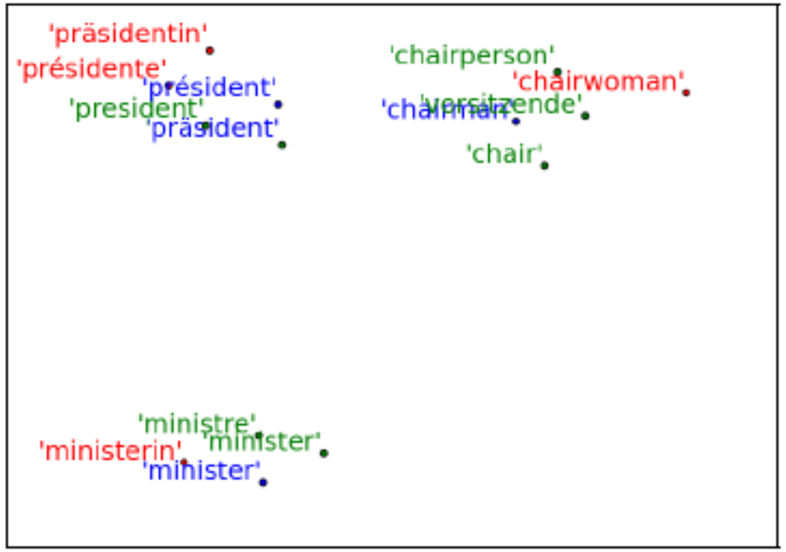
- Classifier trained and evaluated on the same language
- Other embeddings trained on larger datasets
- Performs better with lower amount of data

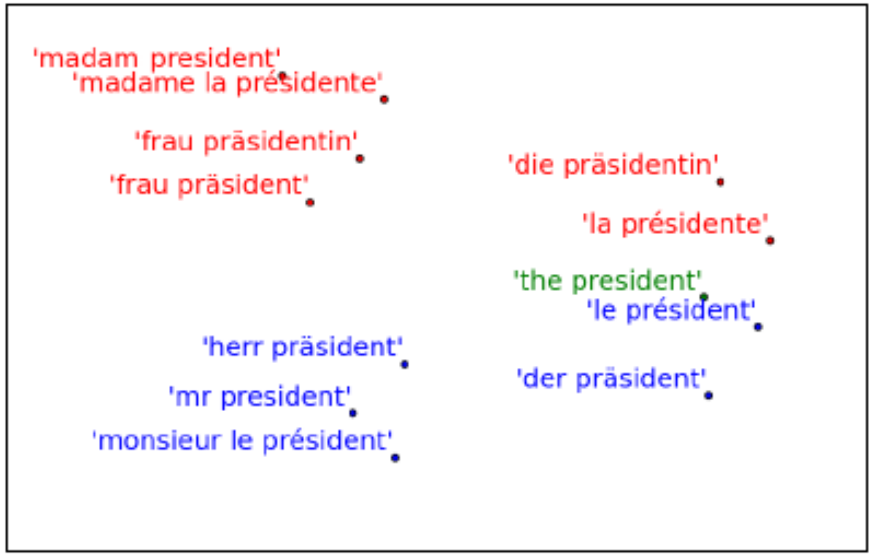
Setting	Languages											
	English	Arabic	German	Spanish	French	Italian	Dutch	Polish	Pt-Br	Roman.	Russian	Turkish
Raw Data NB	0.481	0.469	0.471	0.526	0.532	0.524	0.522	0.415	0.465	0.509	0.465	0.513
Senna	0.400											
Polyglot	0.382	0.416	0.270	0.418	0.361	0.332	0.228	0.323	0.194	0.300	0.402	0.295
single Setting												
DOC/ADD	0.462	0.422	0.429	0.394	0.481	0.458	0.252	0.385	0.363	0.431	0.471	0.435
DOC/BI	0.474	0.432	0.362	0.336	0.444	0.469	0.197	0.414	0.395	0.445	0.436	0.428
joint Setting												
DOC/ADD	0.475	0.371	0.386	0.472	0.451	0.398	0.439	0.304	0.394	0.453	0.402	0.441
DOC/BI	0.378	0.329	0.358	0.472	0.454	0.399	0.409	0.340	0.431	0.379	0.395	0.435

F1-scores for monolingual TED document classification task

Linguistic Analysis

- Linguistic similarity captured even without French-German parallel data
- English serves as pivot
- Separation between the genders





Opinions

- Simple model but possible to extend for advanced embedding models
- Main contribution – loss function and learn without word alignments
- Lot of experiments but high variability in results
- Strictly speaking, is it really compositional?
- No statistical significance but perhaps not presentable

Extensions

- A Multi-Task Approach to Learning Multilingual Representations – Singla et al. (2018)
- Approaches for languages with limited parallel corpora
- Combine with machine translation task
- Scale to more languages
- Advanced composition functions
- Domain adaptation

Questions?

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

Mikel Artetxe
University of the Basque Country (UPV/EHU)*
mikel.artetxe@ehu.eus

Holger Schwenk
Facebook AI Research
schwenk@fb.com

Niels van der Heijden

Creating universal language agnostic sentence embeddings

- Input language agnostic
- NLP task agnostic
- Why?
- Benefit of joint training
- Zero-shot transfer learning
- Code-switching

Short recap of history

Sentence embeddings

- Skip-thought (Kiros et al. 2015)
- NLI (Conneau et al., 2017)
- Multi-task (Cer et al. 2018)

Multilingual representations

- Focused on word embeddings
- Parallel corpora (Gouws et al., 2015)
- Post-processing (Artetxe et al., 2018a)

Short recap of history

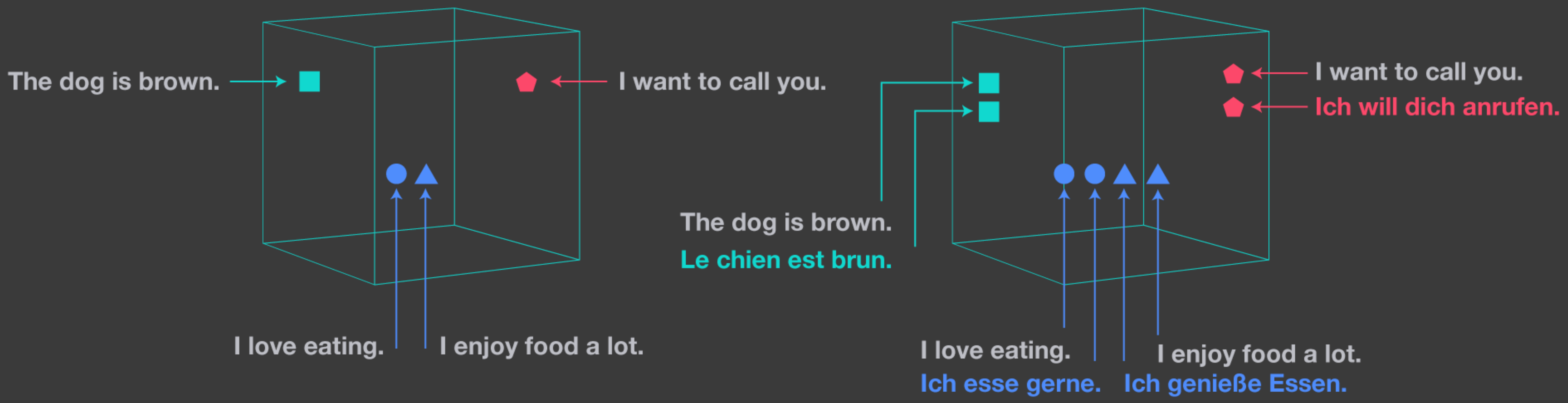
- Seq-to-seq models on parallel corpora (Hassan et al., 2018)
- N-way parallel corpora
- Shared or multiple encoders (Holcher, 2018a)
- Limited number of languages (8)

No work on encoding large amounts of languages into one space

..

Key contributions

- Single sentence encoder for 93 languages
- SOTA on XNLI, BUCC and MLDoc
- New test set for 122 languages
- Training strategy

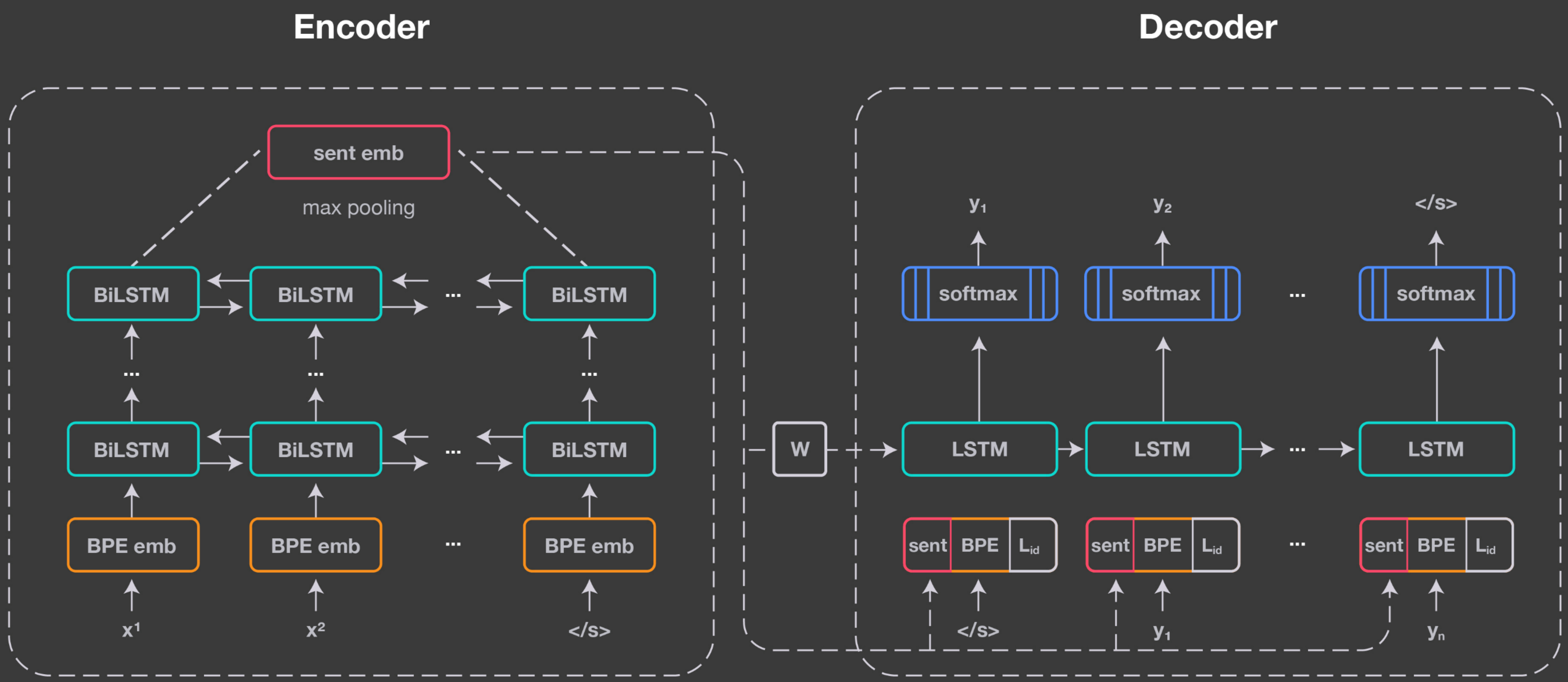


The dog is brown.
I want to call you.
I love eating.
I enjoy food a lot.
I want to call you.
Ich will dich anrufen.
The dog is brown.
Le chien est brun.
I love eating.
Ich esse gerne.
I enjoy food a lot.
Ich genieße Essen.

The data: 223 mln sentences

- Europarl
- United Nations
- OpenSubtitles2018
- Global voices
- Tanzil
- Tatoeba
- http://opus.nlpl.eu

THE MODEL


Encoder
Decoder
sent emb
max pooling
BiLSTM
BiLSTM
BiLSTM
BiLSTM
BiLSTM
BiLSTM
BPE emb
BPE emb
BPE emb
x¹
x²
</s>
W
y₁
y₂
</s>
softmax
softmax
softmax
LSTM
LSTM
LSTM
sent BPE Lid
sent BPE Lid
sent BPE Lid
</s>
y₁
yn

Byte Pair Encoding (Sennrich et al., 2016)

- Need for shared vocabulary
- Words != atomic unit → *Abwasserbehandlungsanlange*
- Iterative procedure
- General purpose

Byte Pair Encoding "collage system"

Text: T = ABABCDEBDEFABDEABC

AB→G

GGCDEBDEFGDEGC

DE→H

GGCHBHFGHGC

GC→I

GIHBHFGHI

D :

$X_1 = \mathrm{A}$;

$X_2 = \mathrm{B}$;

$X_3 = \mathrm{C}$;

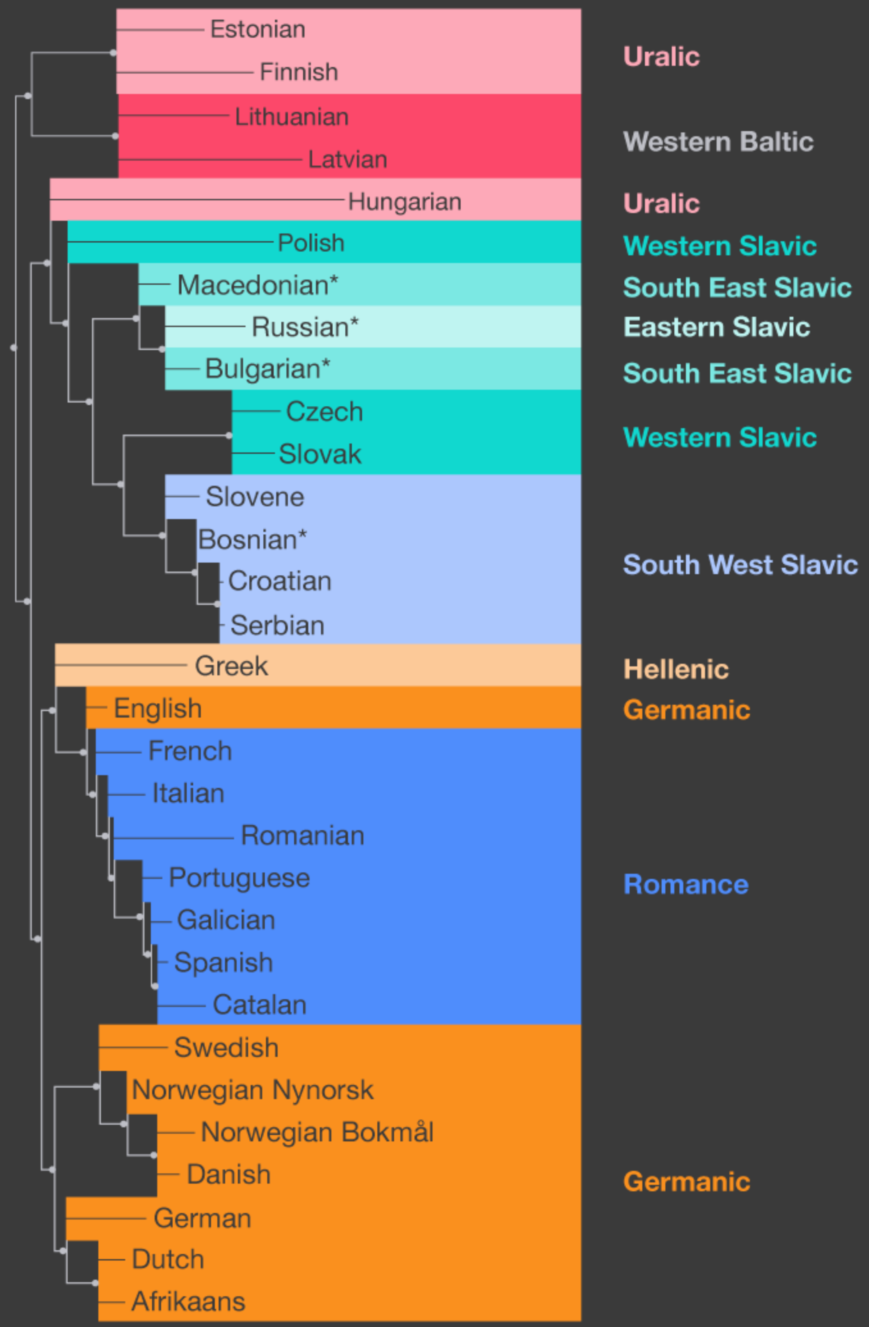
$X_4 = \mathrm{D}$;

$X_5 = \mathrm{E}$;

$X_6 = \mathrm{F}$;

$X_7 = X_1 \cdot X_2$;

$X_8 = X_4 \cdot X_5$;


Estonian
Finnish
Lithuanian
Latvian
Hungarian
Polish
Macedonian*
Russian*
Bulgarian*
Czech
Slovak
Slovene
Bosnian*
Croatian
Serbian
Greek
English
French
Italian
Romanian
Portuguese
Galician
Spanish
Catalan
Swedish
Norwegian Nynorsk
Norwegian Bokmål
Danish
German
Dutch
Afrikaans
Uralic
Western Baltic
Uralic
Western Slavic
South East Slavic
Eastern Slavic
South East Slavic
Western Slavic
South West Slavic
Hellenic
Germanic
Romance
Germanic

Experiments - XNLI

- 15 languages, 2500 dev, 5000 test
- English sentences to 14 languages
- Two layer MLP classifier
- Standard NLI representation (h,|h-p|,h*p,p)

Zero-Shot Transfer, one NLI system for all languages		EN	EN → XX													
			fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
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
(2018c)	X-CBOW	64.5	60.3	60.7	61.0	60.5	60.4	57.8	58.7	57.6	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	74.7	72.3	73.2	72.5	72.7	73.4	71.1	69.8	70.5	71.9	69.2	71.4	66.0	62.1	61.8

Experiments – Tatoeba

- Similarity search
- Cosine similarity
- 122 languages aligned with English
- Test set 1000 sentences per language

Experiments – Tatoeba

amh	am	Amharic	Ethopian	Ge´ez	88k	60.71	55.36	168
ara	ar	Arabic	Arabic	Arabic	8.2M	8.30	7.80	1000
aym	ay	Aymara	Aymaran	Latin	14k	n/a	n/a	–
aze	az	Azerbaijani	Turkic	Latin; Cyrillic; Persian	254k	44.10	23.90	1000
eus	eu	Basque	Isolate	Latin	1.2M	5.70	5.00	1000
ben	bn	Bengali	Indo-Aryan	Eastern-Nagari	913k	10.80	10.00	1000
ber	ber	Berber languages	Berber	Latin	62k	29.80	33.70	1000
nob	nb	Bokmål Norwegian	Germanic	Latin	4.1M	1.30	1.10	1000
bos	bs	Bosnian	Slavic	Latin	4.2M	3.95	3.11	354
bre	br	Breton	Celtic	Latin	29k	83.50	84.90	1000
bul	bg	Bulgarian	Slavic	Cyrillic	4.9M	4.50	5.40	1000
cat	ca	Catalan	Romance	Latin	813k	4.00	4.20	1000
cmn	zh	Chinese mandarin	Chinese	Chinese	8.3M	4.10	5.00	1000
swh	sw	(Coastal) Swahili	Niger-Congo	Latin	173k	45.64	39.23	390
hrv	hr	Croatian	Slavic	Latin	4.0M	2.80	2.70	1000
ces	cs	Czech	Slavic	Latin	5.5M	3.10	3.80	1000
dan	da	Danish	Germanic	Latin	7.9M	3.90	4.00	1000
nld	nl	Dutch	Germanic	Latin	8.4M	3.10	4.30	1000
eng	en	English	Germanic	Latin	2.6M	n/a	n/a	–
epo	eo	Esperanto	constructed	Latin	397k	2.70	2.80	1000
est	et	Estonian	Uralic	Latin	5.3M	3.20	3.40	1000
fin	fi	Finnish	Uralic	Latin	7.9M	3.70	3.70	1000
fra	fr	French	Romance	Latin	8.8M	4.40	4.30	1000
glg	gl	Galician	Romance	Latin	349k	4.60	4.40	1000
kat	ka	Georgian	Kartvelian	Georgian	296k	60.32	67.83	746
deu	de	German	Germanic	Latin	8.7M	0.90	1.00	1000
ell	el	Greek	Hellenic	Greek	6.5M	5.30	4.80	1000
hau	ha	Hausa	Afro-Asiatic	Latin; Arabic	127k	n/a	n/a	–
heb	he	Hebrew	Semitic	Hebrew	4.1M	8.10	7.60	1000
hin	hi	Hindi	Indo-Aryan	Devanagari	288k	5.80	4.80	1000
hun	hu	Hungarian	Uralic	Latin	5.3M	3.90	4.00	1000
isl	is	Icelandic	Germanic	Latin	2.0M	4.40	4.40	1000
ind	id	Indonesian	Malayo-Polynesian	Latin	4.3M	5.20	5.80	1000
pes	ps	Iranian Persian (Farsi)	Iranian	Persian	4.9M	7.20	6.00	1000
ita	it	Italian	Romance	Latin	8.3M	4.60	4.80	1000
jpn	ja	Japanese	Japonic	Kanjii	3.2M	3.90	5.40	1000
kab		Kabyle	Berber	Latin (modified)	15k	39.10	44.70	1000
kor	ko	Korean	Koreanic	Hangul	1.4M	10.60	11.50	1000
kur	ku	Kurdish	Iranian	Latin; Persian	50k	80.24	85.37	410
lvs	lv	Latavian	Baltic	Latin	2.0M	4.50	4.70	1000
lat	la	Latin	Romance	Latin	19k	41.60	41.50	1000
lit	lt	Lithuanian	Baltic	Latin	3.2M	4.10	3.40	1000
nds		Low German / Saxon	Germanic	Latin	12k	18.60	15.60	1000
mkd	mk	Macedonian	Slavic	Cyrillic	4.2M	5.20	5.40	1000
mlg	mg	Malagasy	Malayo-Polynesian	Latin	355k	n/a	n/a	–
zsm	ms	Malay	Malayo-Polynesian	Latin	2.9M	3.40	3.80	1000
mal	ml	Malayalam	Dravidian	Malayalam	373k	3.35	2.91	687
div	dv	Maldivian (Divehi)	Indo-Aryan	Thaana	90k	n/a	n/a	–
mar	mr	Marathi	Indo-Aryan	Devanagari	31k	9.00	8.00	1000
pol	pl	Polish	Slavic	Latin	5.5M	2.00	2.40	1000
por	pt	Portuguese	Romance	Latin	8.3M	4.70	4.90	1000
ron	ro	Romanian; Moldavian	Romance	Latin	4.9M	2.50	2.70	1000
rus	ru	Russian	Slavic	Cyrillic	9.3M	4.90	5.90	1000
srp	sr	Serbian	Slavic	Cyrillic; Latin	4.0M	4.30	5.00	1000
snd	sd	Sindhi	Iranian	Persian; Devanagari	91k	n/a	n/a	–
sin	si	Sinhala	Indo-Aryan	Sinhala	796k	n/a	n/a	–

Experiments – Tatoeba

		Details			Training corpus size	Tatoeba Error [%]		Tatoeba test set size
ISO3	ISO2	Name	Family	Script		en → xx	xx → en	
arq		Algerian Arabic	Arabic	Arabic	none	58.62	62.46	911
ast		Asturian	Romance Ibero	Latin	none	12.60	14.96	127
awa		Awadhi	Indo-Aryan	Devanagari	none	63.20	64.50	231
ceb		Cebuano	Malayo-Polynesian	Latin	none	81.67	87.00	600
cha	ch	Chamorro	Malayo-Polynesian (branch)	Latin	none	64.23	77.37	137
arz		Egyptian Arabic	Arabic	Arabic	none	31.24	31.03	477
fao	fo	Faroese	Germanic	Latin	none	28.24	28.63	262
gla	gd	Gaelic; Scottish Gaelic	Celtic	Latin	none	95.66	96.98	829
jav	jv	Javanese	Malayo-Polynesian	Latin	none	73.66	80.49	205
csb		Kashubian	Slavic	Latin	none	54.55	58.89	253
mon	mn	Mongolian	Mongolic	Cyrillic	none	89.55	94.09	440
max		North Moluccan Malay	Malay Creole	Latin	none	48.24	50.00	284
nov		Novial	constructed	Latin	none	33.07	35.02	257
nno	nn	Nynorsk Norwegian	Germanic	Latin	none	13.40	10.00	1000
ang		Old English	Germanic	Latin	none	58.96	65.67	134
pam		Pampangan; Kapampangan	Philippine	Latin	none	93.10	95.00	1000
pms		Piemontese	Romance	Latin	none	50.86	49.90	525
orv		Russian old	Slavic	Cyrillic	none	68.26	75.45	835
dsb		Sorbian Lower	Slavic	Latin	none	48.64	55.32	479
hsb		Sorbian Upper	Slavic	Latin	none	42.44	48.65	483
swg		Swabian	Germanic	Latin	none	50.00	58.04	112
gsw		Swiss German	Germanic	Latin	none	52.99	58.12	117
tzl		Talossan	constructed	Latin	none	54.81	55.77	104
tuk	tk	Turkmen	Turkic	Latin	none	75.37	83.25	203
war		Waray	Malayo-Polynesian	Latin	none	84.20	88.60	1000
cym	cy	Welsh	Celtic	Latin-Welsch	none	89.74	93.04	575
fry	fy	Western Frisian	Germanic	Latin	none	46.24	50.29	173
xho	xh	Xhosa	Niger-Congo	Latin	none	90.85	92.25	142
yid	yi	Yiddish	Germanic	Hebrew	none	93.28	95.40	848

- 48 < 10%
- 55 < 20%
- 15 > 50%
- But: performance can still be good on completely unseen languages

Conclusion

- New SOTA on XNLI, MLDoc and BUCC
- New Tatoeba test set

Opinion

- Overall well written
- Two completely novel contributions
- Elaborate appendix & ablation experiments

- Little analysis on generalization gap between XNLI and Tatoeba
- Train data not open-sourced
- No significance anywhere
- Simple BiLSTM encoder a bit naïve

- Replace LSTM encoder
- Word-level capabilities

Future research

	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**

XLM BERT (LAMPLE ET AL., 2019)

THANK YOU