Multilingual Sentence Embeddings in/and/for Neural Machine Translation

Cristina España-Bonet
DFKI GmbH

Recent Advances in Machine Translation (RAMT 2021)

Webex, everywhere on the Earth
(with internet)
18th March 2021

What’s all this about?

NIT-Silchar

UdS

Multimodal Machine Translation, Convergence of Multiple Input Modes

National Coordinator

SPARC

Scheme for Promotion of Academic and Research Collaboration

What's all this about?

Neural Machine Translation (NMT)
text2text

Self-Supervised NMT

What's all this about?

RAMT: Recent Advances in Machine Translation

Neural Machine Translation (NMT)
text2text

Self-Supervised NMT

Multi/Cross-lingual Embeddings

What's all this about?

RAMT: Recent Advances in Machine Translation

Neural Machine Translation (NMT)
text2text

Self-Supervised NMT

Multi/Cross-lingual Embeddings

NMT Initialisation

What's all this about?

Relations with

Josef van Genabith tutorial on NMT (Monday)

Mikel Artetxe talk on Unsupervised NMT (Tomorrow)

What’s all this about?

Relations with

Josef van Genabith tutorial on NMT (Monday)

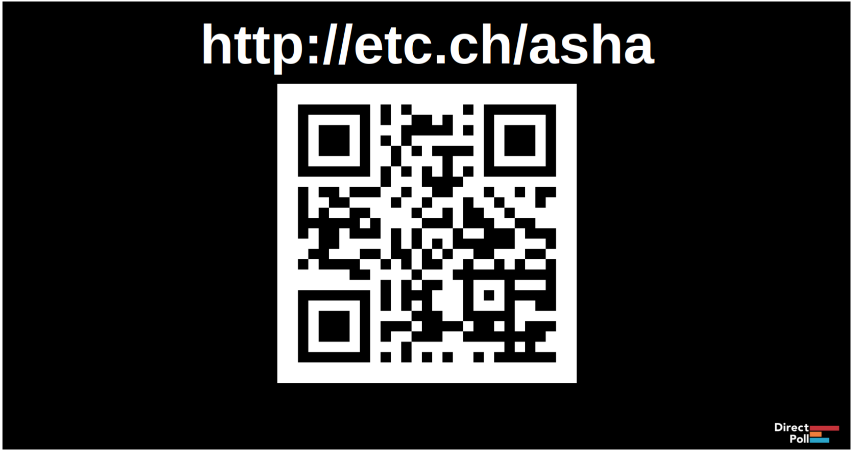
Mikel Artetxe talk on Unsupervised NMT (Tomorrow)

Let’s go interactive!

`https://directpoll.com/r?XDbzPBd3ixYqg8eeRn4nQFkQZJV3t8WBbAqGR5Y7f`

What's all this about?

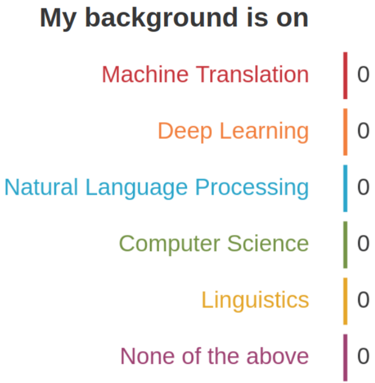
Let's go interactive! DirectPoll



http://etc.ch/asha

What's all this about?

Let's go interactive! DirectPoll

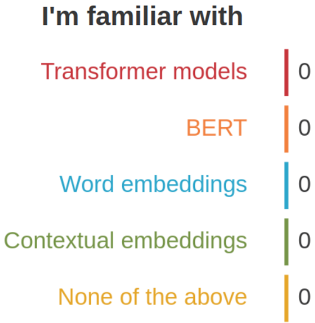


My background is on

Machine Translation	0
Deep Learning	0
Natural Language Processing	0
Computer Science	0
Linguistics	0
None of the above	0

What's all this about?

Let's go interactive! DirectPoll



I'm familiar with

Transformer models	0
BERT	0
Word embeddings	0
Contextual embeddings	0
None of the above	0

Outline

1. Motivation

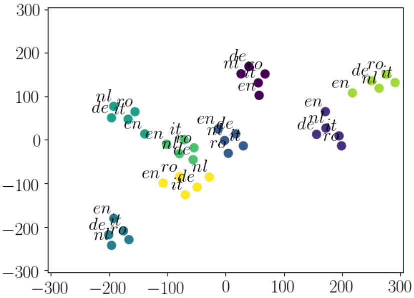
2. (Multilingual) Sentence Embeddings

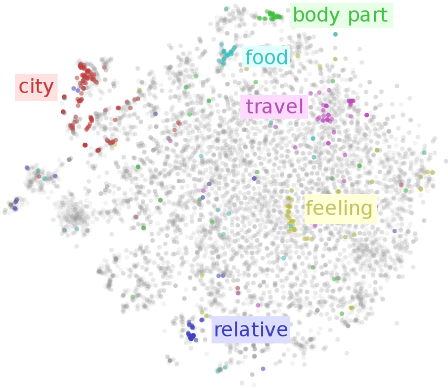
3. Self-Supervised NMT

4. Initialising (Multilingual) NMT

Basic Concepts (Josef's Tutorial)

Static/Contextual/Sentence Embeddings





https://ruder.io/word-embeddings-1/

Basic Concepts (Josef's Tutorial)

NMT with Transformers

embeddings ⇝

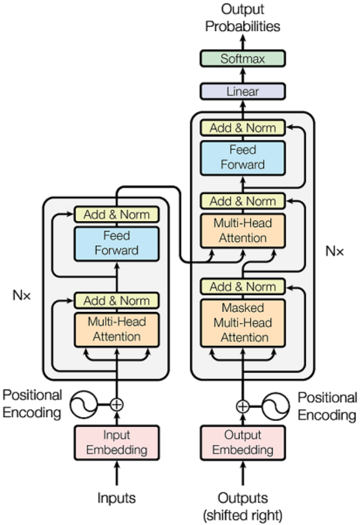
embeddings ⇝

embeddings ⇝

embeddings ⇝

embeddings ⇝

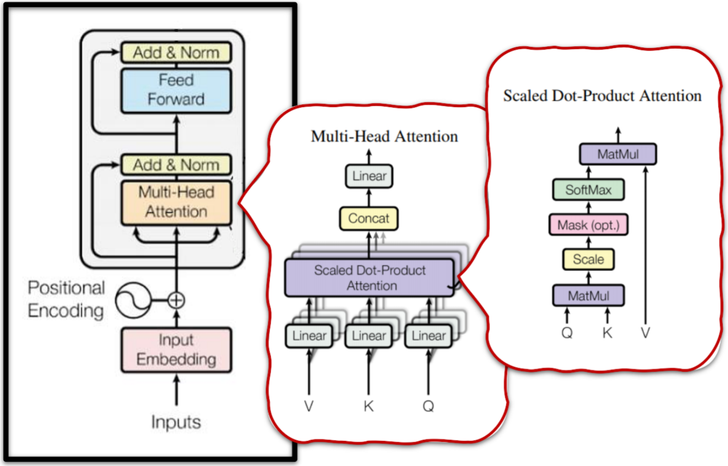
embeddings ⇝



(Vaswani et al., 2017)

Basic Concepts (Josef’s Tutorial)

NLP 2020 Summary: Transformer Blocks



(Multilingual) Sentence Embeddings

Semantic Similarity and Parallel Sentences

- This is presentation is about **machine translation**
  - by definition a **multilingual** (bilingual) task
  - translations are cross-lingual pairs of sentences with **similarity** 1
- Lot of work on semantic similarity between embeddings
- Can **multilingual embeddings** be a good tool here?
  - ✓ for parallel sentence selection
  - ✓ for initialisation (word/sentence embeddings)
- What is a good representation of a sentence?

(Multilingual) Sentence Embeddings

Sentence Embeddings (keywords to google after the talk)

- Averaging (weighting) word embeddings
- Sent2Vec / Paragraph vectors (doc2vec) / Doc2VecC
- Skip-thought / FastSent / Quick-thought vectors
- Sentence-BERT (SBERT) / LASER / T-LASER / GPT, ...
- Averaging (weighting) sentence embeddings for document embedding

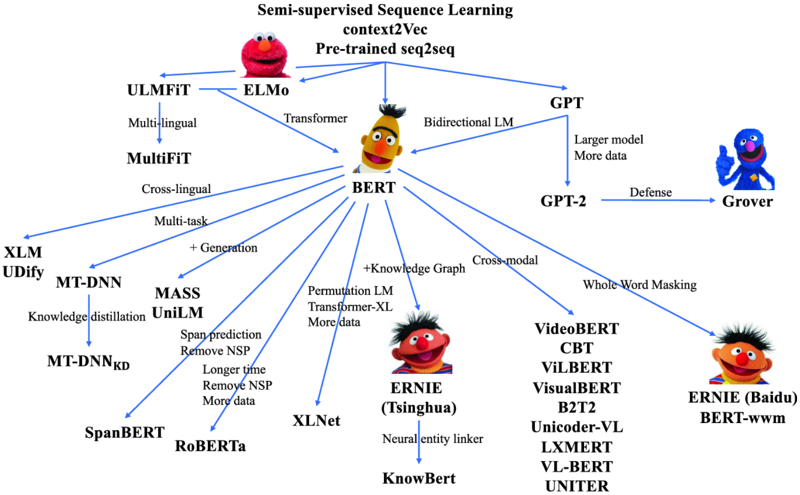
(Multilingual) Sentence Embeddings

Word vs. Sentence Embeddings

- Word embeddings are basic units in NLP
- Contextualised (BERT-like) embeddings
  - solve ambiguity problems of static (word2vec-like) embeddings
  - include a “sentence representation” token ([CLS])
  - are easily and successfully fine-tuned to several NLP tasks
  - without fine tuning, performance drops
- Lots of sentence embeddings, I start with BERT because its common usage and number of *relatives*

(Multilingual) Sentence Embeddings

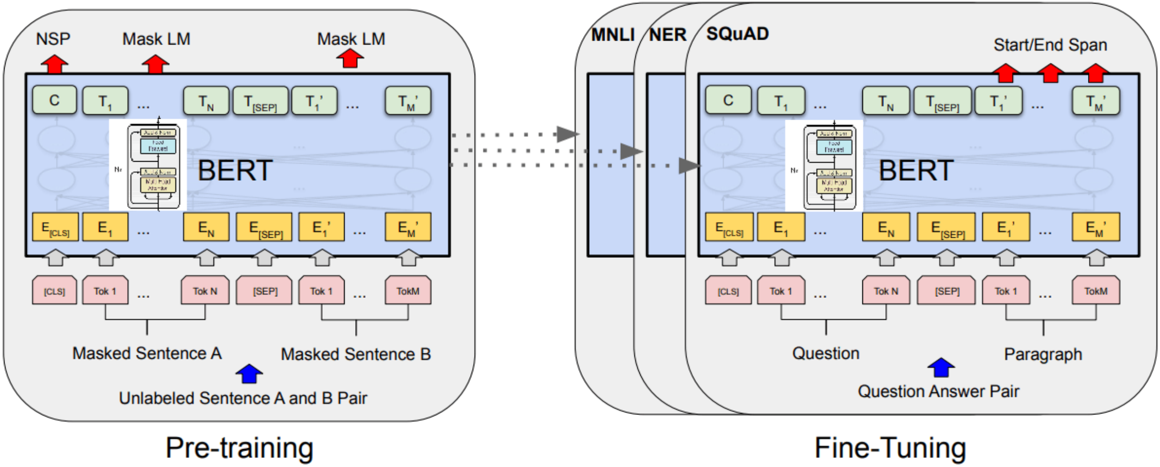
BERT Relatives



(Liu et al, 2020)

(Multilingual) Sentence Embeddings with BERT

BERT Model: stack of TF blocks train for NSP and Mask LM



(Multilingual) Sentence Embeddings with BERT

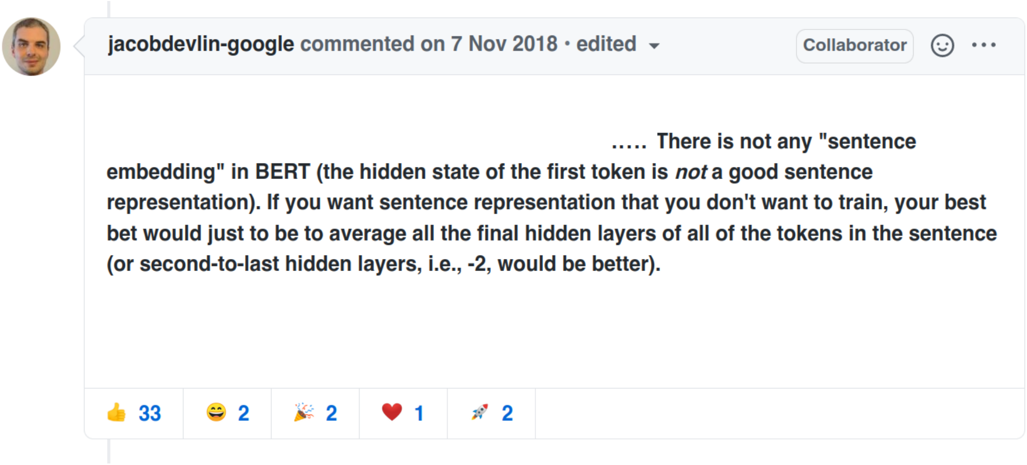
BERT Applications

Everything and more. But designed for fine-tuning on:

- Sentence classification tasks
  - `[CLS] An individual sentence goes here`
- Sentence-pair regression tasks
  - `[CLS] Sentence one here [SEP] Sentence 2 after the first one`

(Multilingual) Sentence Embeddings with BERT

BERT Non-Applications: Sentence Embeddings (without FT)



jacobdevlin-google commented on 7 Nov 2018 • edited

..... There is not any "sentence embedding" in BERT (the hidden state of the first token is *not* a good sentence representation). If you want sentence representation that you don't want to train, your best bet would just to be to average all the final hidden layers of all of the tokens in the sentence (or second-to-last hidden layers, i.e., -2, would be better).

👍 33 😄 2 🎉 2 ❤️ 1 🚀 2

(https://github.com/google-research/bert/issues/71)

Digression

Semantic Textual Similarity (STS)

Digression

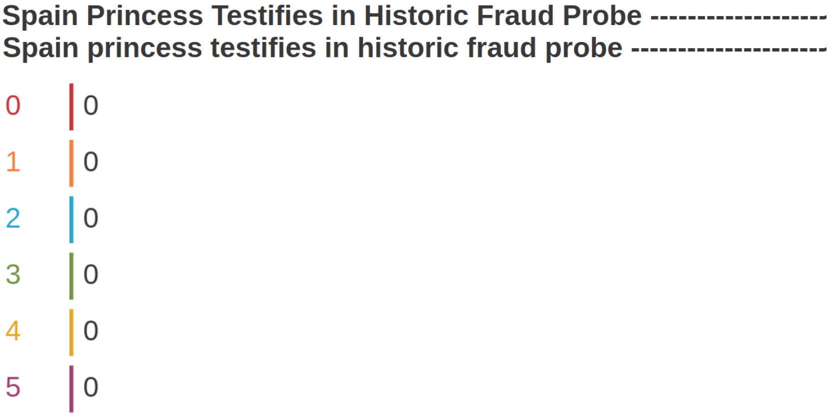
Semantic Textual Similarity (STS)

STS measures the degree of equivalence in the underlying semantics of paired snippets of text

“Given two sentences, the task is to return a **continuous valued similarity score on a scale from 0 to 5**, with 0 indicating that the semantics of the sentences are completely independent and 5 signifying semantic equivalence.”

Digression

STS Example (DirectPoll)

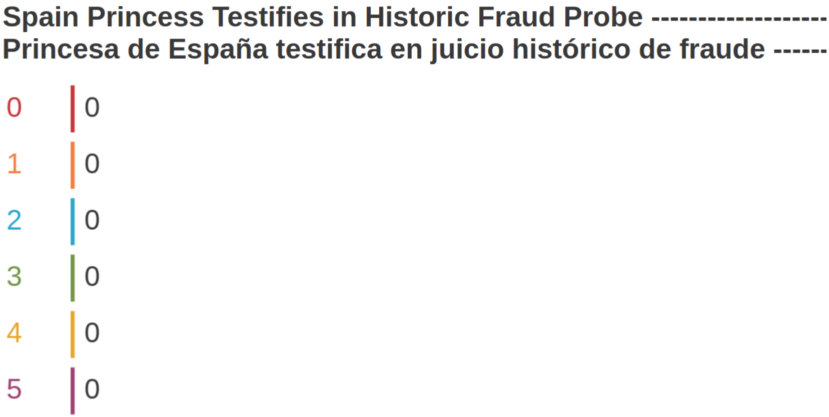


Spain Princess Testifies in Historic Fraud Probe ------------------
Spain princess testifies in historic fraud probe -------------------

0	0
1	0
2	0
3	0
4	0
5	0

Digression

STS Example (DirectPoll)

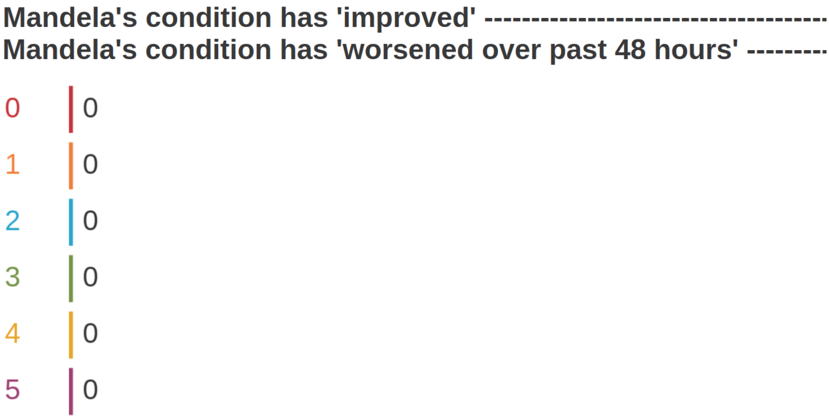


Spain Princess Testifies in Historic Fraud Probe ------------------
Princesa de España testifica en juicio histórico de fraude ------

0	0
1	0
2	0
3	0
4	0
5	0

Digression

STS Example (DirectPoll)



Mandela's condition has 'improved' ------------------------------------
Mandela's condition has 'worsened over past 48 hours' ---------

0	0
1	0
2	0
3	0
4	0
5	0

Digression

Semantic Textual Similarity (STS)

STS measures the degree of equivalence in the underlying semantics of paired snippets of text

“Given two sentences, the task is to return a continuous valued similarity score on a scale from 0 to 5, with 0 indicating that the semantics of the sentences are completely independent and 5 signifying semantic equivalence.”

Evaluation: Pearson correlation or Spearman’s rank **correlation** between the cosine similarity of the sentence embeddings and the gold labels

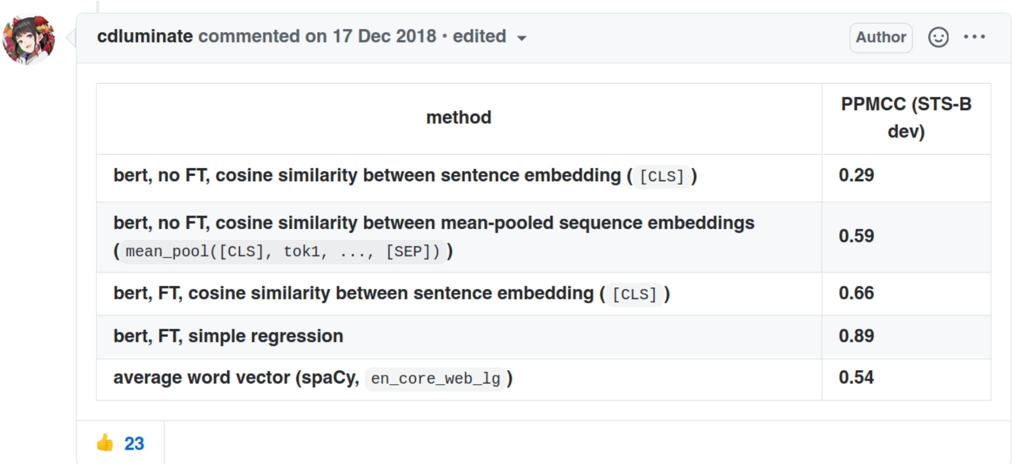
Digression

Let's join the main path again

BERT on STS

(Multilingual) Sentence Embeddings with BERT

BERT Sentence Embeddings on STS



cdluminate commented on 17 Dec 2018 • edited

method	PPMCC (STS-B dev)
bert, no FT, cosine similarity between sentence embedding (`[CLS]`)	0.29
bert, no FT, cosine similarity between mean-pooled sequence embeddings (`mean_pool([CLS], tok1, ..., [SEP])`)	0.59
bert, FT, cosine similarity between sentence embedding (`[CLS]`)	0.66
bert, FT, simple regression	0.89
average word vector (spaCy, `en_core_web_lg`)	0.54

👍 23

(https://github.com/google-research/bert/issues/276)

(Multilingual) Sentence Embeddings with BERT

BERT Sentence Embeddings on STS

Pearson correlation on STS 2017 data

	track1	track2	track3	track4a	track5
	ar–ar	*ar–en*	*es–es*	*es–en*	*en–en*
WE-d300	0.49	0.28	0.55	0.40	0.56
WE-d1024	0.51	0.33	0.59	0.45	0.60
NMT$_{\mathrm{ctx}}$-2.0Ep	0.59	0.44	0.78	0.49	0.76
BERT	?	?	?	?	*0.59*
BERT+FT	?	?	?	?	0.85
BERT$_{LARGE}$+FT	?	?	?	?	0.86

(España-Bonet et al., 2017)

(Multilingual) Sentence Embeddings with BERT

BERT Sentence Embeddings on STS, no FT

Spearman rank correlation on several STS sets

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICKR	Avg.
Avg. GloVe emb.	0.55	0.71	0.60	0.68	0.64	0.58	0.54	0.61
Avg. BERT emb.	0.39	0.58	0.58	0.63	0.61	0.46	0.58	0.55
BERT CLS-vec	0.20	0.30	0.20	0.37	0.38	0.16	0.43	0.29

(Reimers and Gurevych, 2019)

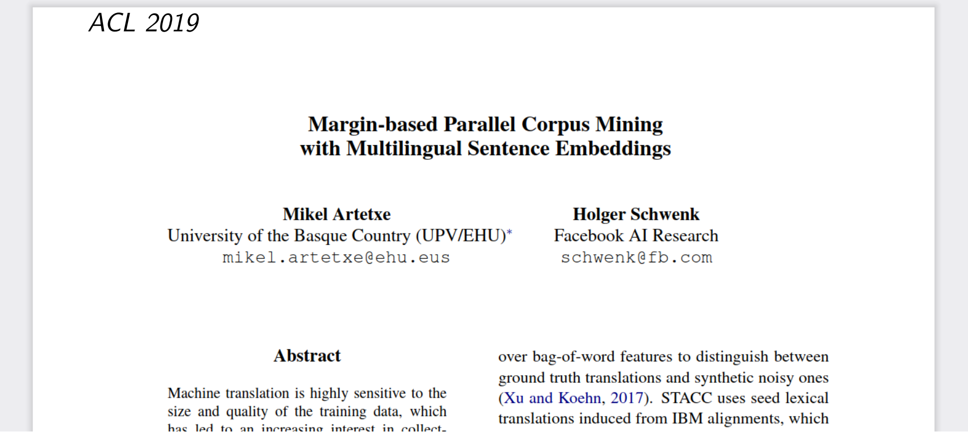
(Multilingual) Sentence Embeddings

Remember our Questions

- This is presentation is about machine translation
  - by definition a multilingual (bilingual) task
  - translations are cross-lingual pairs of sentences with similarity 1
- **What is a good representation of a sentence?**
- **Can multilingual embeddings be a good tool here?**
  - **for parallel sentence selection**
  - for initialisation (word/sentence embeddings)

Multilingual Sentence Embeddings with LASER

Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings



ACL 2019

Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

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

Holger Schwenk
Facebook AI Research
schwenk@fb.com

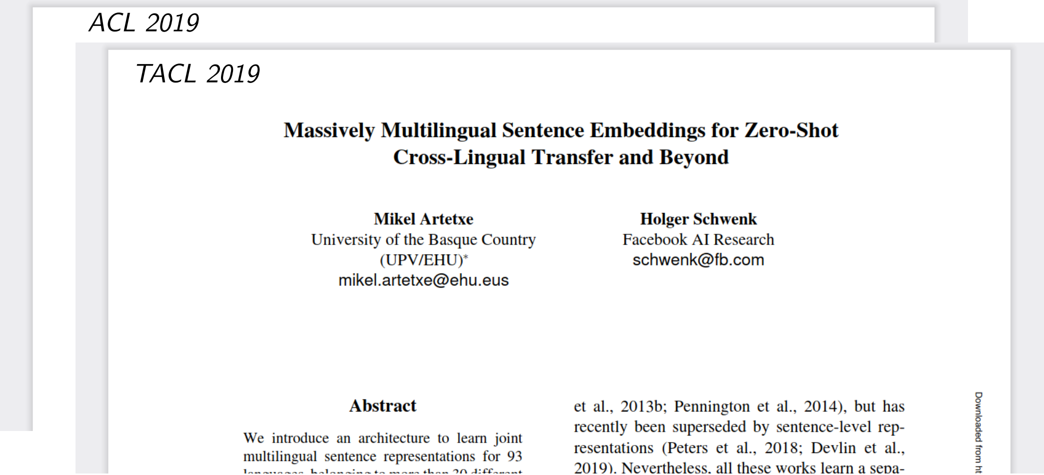
Abstract

Machine translation is highly sensitive to the size and quality of the training data, which has led to an increasing interest in collect-

over bag-of-word features to distinguish between ground truth translations and synthetic noisy ones (Xu and Koehn, 2017). STACC uses seed lexical translations induced from IBM alignments, which

Multilingual Sentence Embeddings with LASER

Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings



ACL 2019

TACL 2019

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

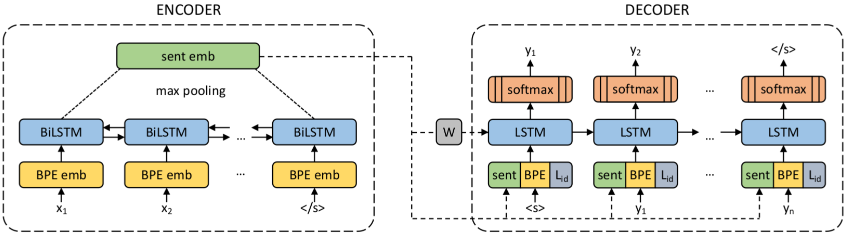
Abstract

We introduce an architecture to learn joint multilingual sentence representations for 93 languages, belonging to more than 30 different

et al., 2013b; Pennington et al., 2014), but has recently been superseded by sentence-level representations (Peters et al., 2018; Devlin et al., 2019). Nevertheless, all these works learn a sepa-

Multilingual Sentence Embeddings with LASER

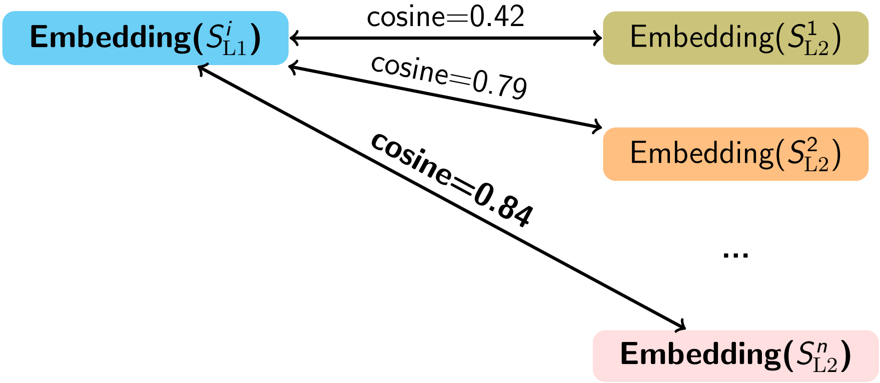
Architecture (based on Schwenk 2018)



- Training with (multilingual) parallel corpora, MT task
- Sentence embeddings from the language agnostic encoder
- **L**anguage **A**gnostic **SE**ntence **R**epresentations: 1024-dim embeddings

Multilingual Sentence Embeddings with LASER

The Key Point: Margin-based Similarity for Scoring Pairs



Embedding(S^{i}_{L1})

cosine=0.42

Embedding(S^{1}_{L2})

cosine=0.79

Embedding(S^{2}_{L2})

cosine=0.84

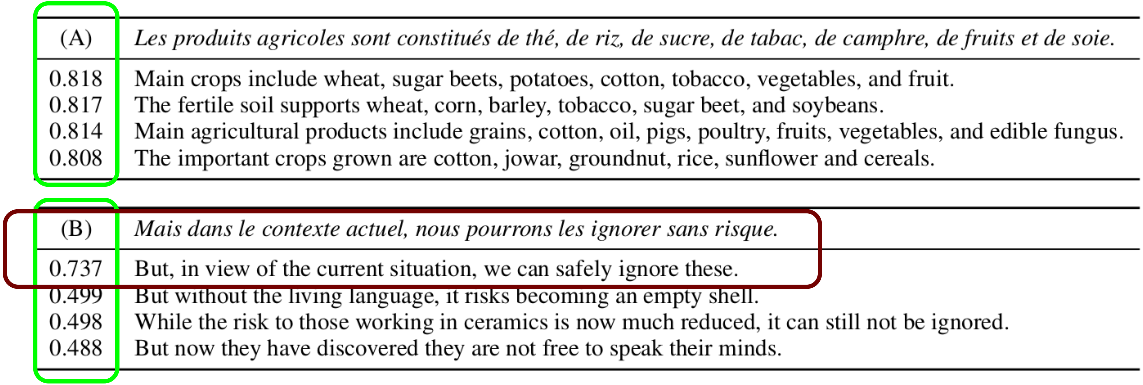
...

Embedding(S^{n}_{L2})

Threshold=0.80 ($\forall i$)

Multilingual Sentence Embeddings with LASER

The Key Point: Margin-based Similarity for Scoring Pairs



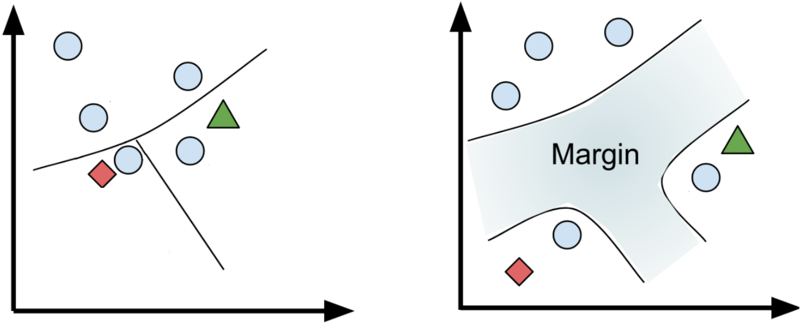
(A)	*Les produits agricoles sont constitués de thé, de riz, de sucre, de tabac, de camphre, de fruits et de soie.*
0.818	Main crops include wheat, sugar beets, potatoes, cotton, tobacco, vegetables, and fruit.
0.817	The fertile soil supports wheat, corn, barley, tobacco, sugar beet, and soybeans.
0.814	Main agricultural products include grains, cotton, oil, pigs, poultry, fruits, vegetables, and edible fungus.
0.808	The important crops grown are cotton, jowar, groundnut, rice, sunflower and cereals.

(B)	*Mais dans le contexte actuel, nous pourrons les ignorer sans risque.*
0.737	But, in view of the current situation, we can safely ignore these.
0.499	But without the living language, it risks becoming an empty shell.
0.498	While the risk to those working in ceramics is now much reduced, it can still not be ignored.
0.488	But now they have discovered they are not free to speak their minds.

- Cosine similarity has a different scale per sentence

Multilingual Sentence Embeddings with LASER

The Key Point: Margin-based Similarity for Scoring Pairs



Cosine accepted pairs

Margin accepted pairs

(Adapted from Yang et al, 2019)

Multilingual Sentence Embeddings with LASER

The Key Point: Margin-based Similarity for Scoring Pairs

$$\mathrm{margin}_{\mathrm{LASER}}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \frac{\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})}{\mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 + \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2}$$

where $\mathrm{avr}_{\mathrm{kNN}}(X, Y_k) = \sum_{Y \in k\mathrm{NN}(X)} \frac{\cos(X,Y)}{k}$ (average similarity)

Multilingual Sentence Embeddings with LASER

The Key Point: Margin-based Similarity for Scoring Pairs

Artetxe et al.

$$\mathrm{margin}_{\mathrm{LASER}}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \frac{\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})}{\mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 + \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2}$$

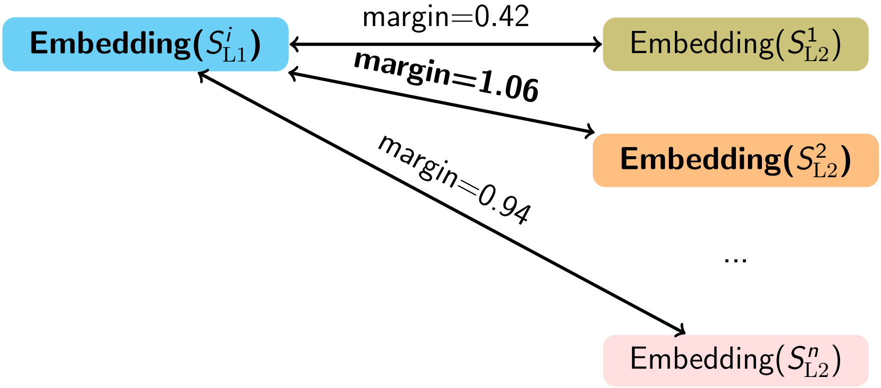
Conneau et al., 2018

$$\mathrm{margin}_{\mathrm{CSLS}}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \cos(S_{\mathrm{L1}}, S_{\mathrm{L2}}) - \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 - \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2$$

where $\mathrm{avr}_{\mathrm{kNN}}(X, Y_k) = \sum_{Y \in k\mathrm{NN}(X)} \frac{\cos(X,Y)}{k}$ (average similarity)

Multilingual Sentence Embeddings with LASER

The Key Point: Margin-based Similarity for Scoring Pairs



Embedding(S^{i}_{L1})

margin=0.42

Embedding(S^{1}_{L2})

margin=1.06

Embedding(S^{2}_{L2})

margin=0.94

...

Embedding(S^{n}_{L2})

Threshold=1.04 **($\forall i$)**

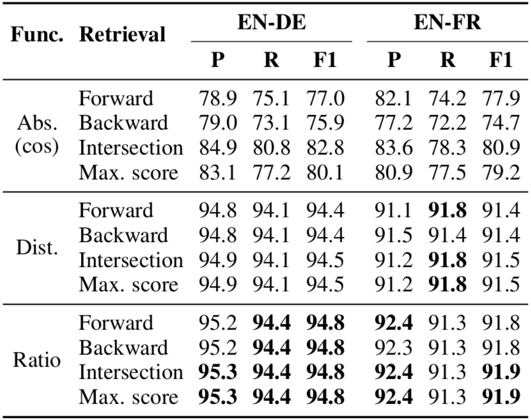
Multilingual Sentence Embeddings with LASER

Parallel Sentence Extraction

$$\cos(S_{\mathrm{L}1}, S_{\mathrm{L}2})$$

$$\mathrm{margin}_{\mathrm{CSLS}}(S_{\mathrm{L}1}, S_{\mathrm{L}2})$$

$$\mathrm{margin}_{\mathrm{LASER}}(S_{\mathrm{L}1}, S_{\mathrm{L}2})$$



Func.	**Retrieval**	**EN-DE**			**EN-FR**		
		P	**R**	**F1**	**P**	**R**	**F1**
Abs. (cos)	Forward	78.9	75.1	77.0	82.1	74.2	77.9
	Backward	79.0	73.1	75.9	77.2	72.2	74.7
	Intersection	84.9	80.8	82.8	83.6	78.3	80.9
	Max. score	83.1	77.2	80.1	80.9	77.5	79.2
Dist.	Forward	94.8	94.1	94.4	91.1	**91.8**	91.4
	Backward	94.8	94.1	94.4	91.5	91.4	91.4
	Intersection	94.9	94.1	94.5	91.2	**91.8**	91.5
	Max. score	94.9	94.1	94.5	91.2	**91.8**	91.5
Ratio	Forward	95.2	**94.4**	**94.8**	**92.4**	91.3	91.8
	Backward	95.2	**94.4**	**94.8**	92.3	91.3	91.8
	Intersection	**95.3**	**94.4**	**94.8**	**92.4**	91.3	**91.9**
	Max. score	**95.3**	**94.4**	**94.8**	**92.4**	91.3	**91.9**

Table 2: BUCC results (precision, recall and F1) on the training set, used to optimize the filtering threshold.

Multilingual Sentence Embeddings with LASER

Applications

Mining of parallel corpora

- **WikiMatrix**: Mining 135M Parallel Sent. in 1620 Language Pairs from WP
- **CCMatrix**: Mining Billions of High-Quality Parallel Sentences on the WEB
- `https://github.com/facebookresearch/LASER`

Multilingual Sentence Embeddings with LASER

Applications

Mining of parallel corpora

- **WikiMatrix**: Mining 135M Parallel Sent. in 1620 Language Pairs from WP
- **CCMatrix**: Mining Billions of High-Quality Parallel Sentences on the WEB
- `https://github.com/facebookresearch/LASER`

Others

- Cross-lingual Natural Language Inference (XNLI)
- Cross-lingual text classification
- Cross-lingual similarity search

Multilingual Sentence Embeddings with LASER

Limitations and Enhancements

- Great for bitext identification ($sim = 5$), even zero-shot
- Weaker for semantic similarity tasks ($0 < sim < 5$) —see later
  - Common trend for systems trained on the MT task alone

Multilingual Sentence Embeddings with LASER

Limitations and Enhancements

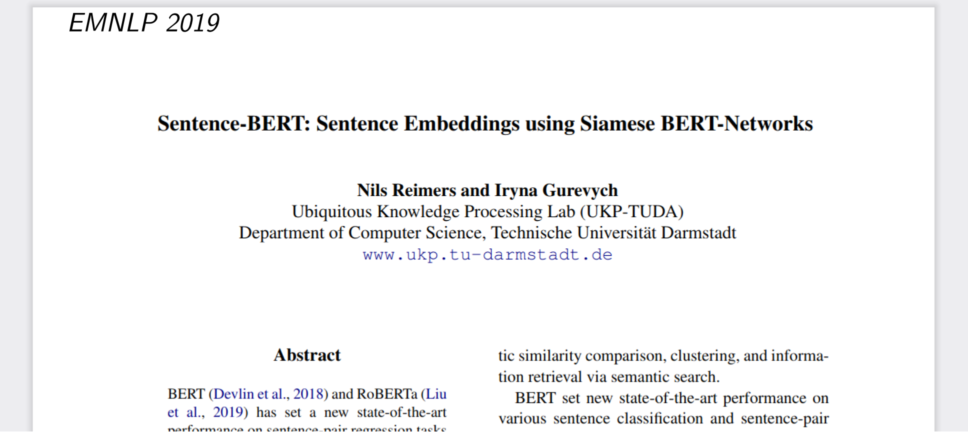
- Great for bitext identification ($sim = 5$), even zero-shot
- Weaker for semantic similarity tasks ($0 < sim < 5$) —see later
    - Common trend for systems trained on the MT task alone
- Version with a Transformer encoder instead of the BiLSTM and modification of the loss function in LASER-cT

  Transformer based Multilingual document Embedding model

  Wei Li, Brian Mak (2020)
    - no pre-trained multilingual version :-(

Multilingual Sentence Embeddings with SBERT and MKD

Making Monolingual Sentence Embeddings ML using Knowledge Distillation



EMNLP 2019

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

Nils Reimers and Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Abstract

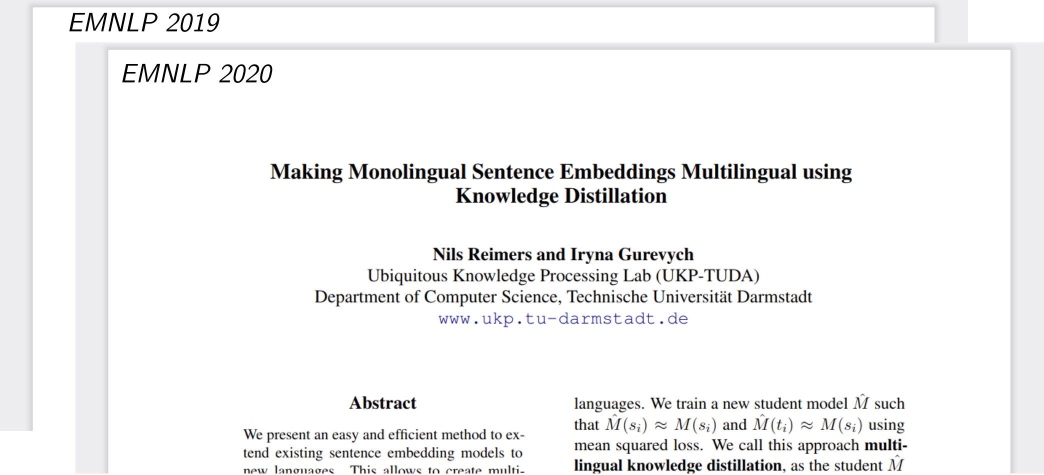
BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) has set a new state-of-the-art performance on sentence-pair regression tasks

tic similarity comparison, clustering, and information retrieval via semantic search.

BERT set new state-of-the-art performance on various sentence classification and sentence-pair

Multilingual Sentence Embeddings with SBERT and MKD

Making Monolingual Sentence Embeddings ML using Knowledge Distillation



EMNLP 2019

EMNLP 2020

Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation

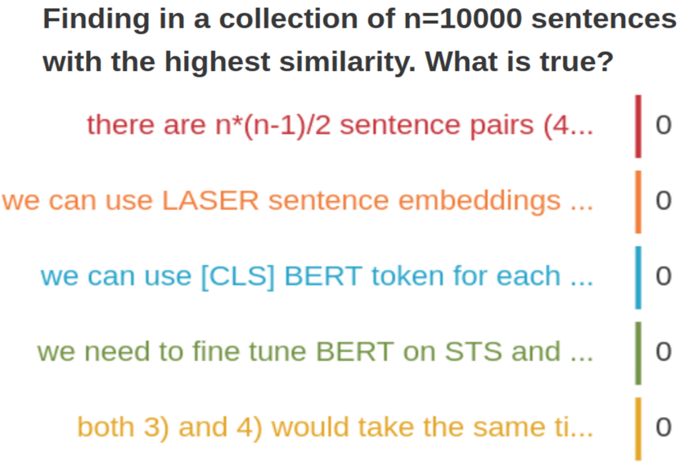
Nils Reimers and Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Abstract

We present an easy and efficient method to extend existing sentence embedding models to new languages. This allows to create multi- languages. We train a new student model $\hat{M}$ such that $\hat{M}(s_i) \approx M(s_i)$ and $\hat{M}(t_i) \approx M(s_i)$ using mean squared loss. We call this approach **multilingual knowledge distillation**, as the student $\hat{M}$

Multilingual Sentence Embeddings with SBERT and MKD

Work Motivation, can you Guess? (DirectPoll)

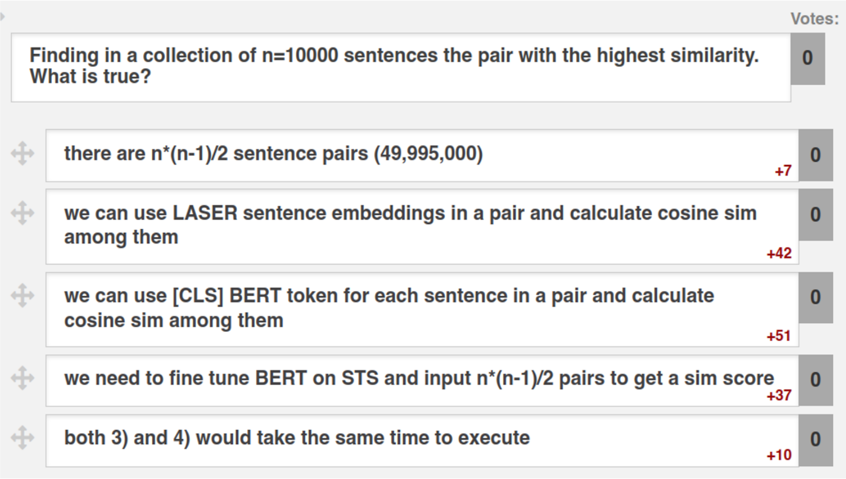


Finding in a collection of n=10000 sentences the pair with the highest similarity. What is true?

there are n*(n-1)/2 sentence pairs (4...	0
we can use LASER sentence embeddings ...	0
we can use [CLS] BERT token for each ...	0
we need to fine tune BERT on STS and ...	0
both 3) and 4) would take the same ti...	0

Multilingual Sentence Embeddings with SBERT and MKD

Work Motivation, can you Guess? (DirectPoll)



Votes:

Finding in a collection of n=10000 sentences the pair with the highest similarity. What is true? 0

there are n*(n-1)/2 sentence pairs (49,995,000) +7 0

we can use LASER sentence embeddings in a pair and calculate cosine sim among them +42 0

we can use [CLS] BERT token for each sentence in a pair and calculate cosine sim among them +51 0

we need to fine tune BERT on STS and input n*(n-1)/2 pairs to get a sim score +37 0

both 3) and 4) would take the same time to execute +10 0

Multilingual Sentence Embeddings with SBERT and MKD

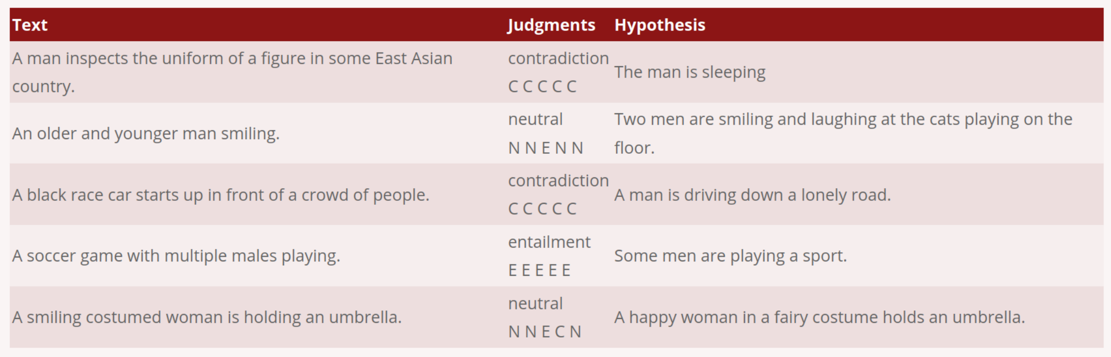
Sentence-BERT (SBERT)

- SBERT adds a pooling operation to the output of BERT
- Fine-tune with NLI data of a
  - Siamese network
  - triplet network (Siamese with triplet objective function)
- NLI data have been shown to be the best for general sentence embeddings

Multilingual Sentence Embeddings with SBERT and MKD

Brief Background: NLI Data

Example from SNLI dataset

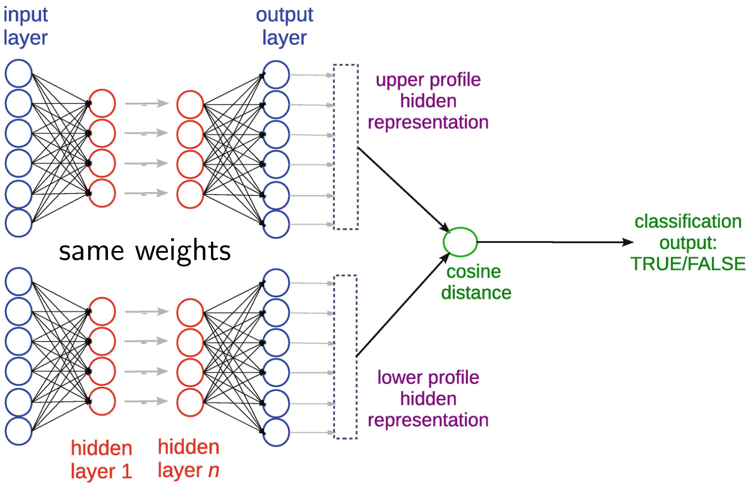


Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

https://nlp.stanford.edu/projects/snli/

Multilingual Sentence Embeddings with SBERT and MKD

Brief Background: Siamese Neural Networks


input layer
output layer
upper profile hidden representation
same weights
cosine distance
classification output: TRUE/FALSE
lower profile hidden representation
hidden layer 1
hidden layer *n*

(https://link.springer.com/protocol/10.1007/978-1-0716-0826-5_3)

Multilingual Sentence Embeddings with SBERT and MKD

SBERT Architecture

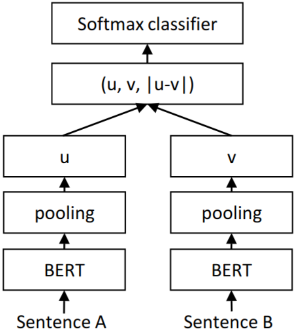


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

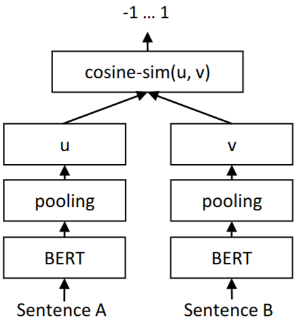
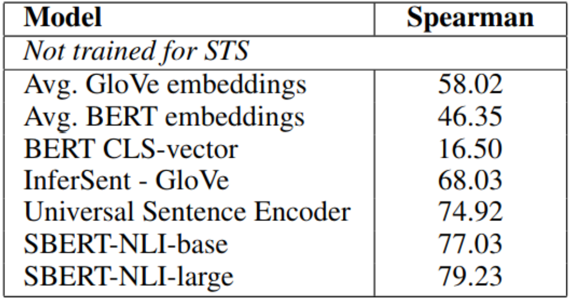


Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

Multilingual Sentence Embeddings with SBERT and MKD

SBERT Results on STSb (Unsupervised)

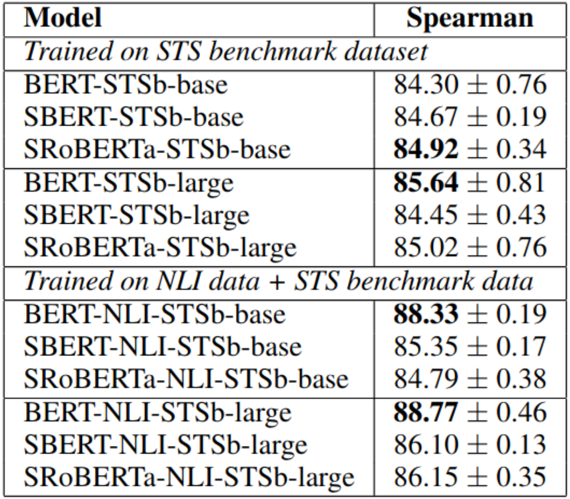


Model	**Spearman**
Not trained for STS	
Avg. GloVe embeddings	58.02
Avg. BERT embeddings	46.35
BERT CLS-vector	16.50
InferSent - GloVe	68.03
Universal Sentence Encoder	74.92
SBERT-NLI-base	77.03
SBERT-NLI-large	79.23

- Remember the difficulty of manualy scoring pairs for similarity
- Correlation of 80 is good!

Multilingual Sentence Embeddings with SBERT and MKD

SBERT Results on STSb (Supervised)



Model	**Spearman**
Trained on STS benchmark dataset	
BERT-STSb-base	84.30 ± 0.76
SBERT-STSb-base	84.67 ± 0.19
SRoBERTa-STSb-base	**84.92** ± 0.34
BERT-STSb-large	**85.64** ± 0.81
SBERT-STSb-large	84.45 ± 0.43
SRoBERTa-STSb-large	85.02 ± 0.76
Trained on NLI data + STS benchmark data	
BERT-NLI-STSb-base	**88.33** ± 0.19
SBERT-NLI-STSb-base	85.35 ± 0.17
SRoBERTa-NLI-STSb-base	84.79 ± 0.38
BERT-NLI-STSb-large	**88.77** ± 0.46
SBERT-NLI-STSb-large	86.10 ± 0.13
SRoBERTa-NLI-STSb-large	86.15 ± 0.35

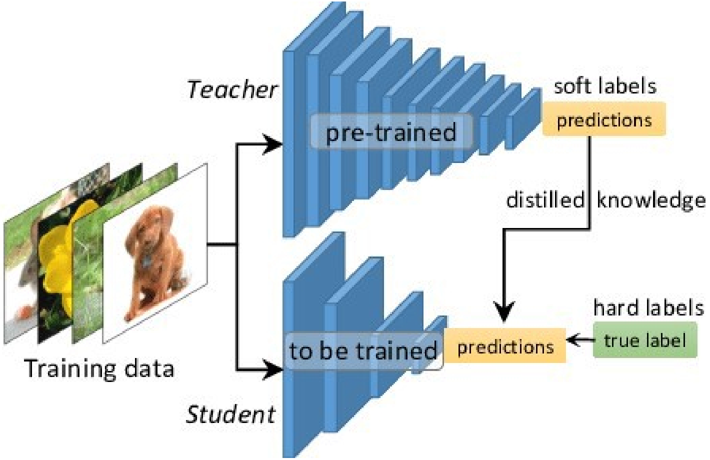
Multilingual Sentence Embeddings with SBERT and MKD

Multilingual Knowledge Distillation

We have monolingual sentence embeddings.
Now what?

Multilingual Sentence Embeddings with SBERT and MKD

Brief Background: Knowledge Distillation



Multilingual Sentence Embeddings with SBERT and MKD

Multilingual Knowledge Distillation

Idea

Monolingual Sentence Embeddings L1

Parallel corpus L1–L2, with sentences s_i^{L1}, t_i^{L2} (or more languages)

Multilingual Sentence Embeddings with SBERT and MKD

Multilingual Knowledge Distillation

Idea

(Good) Monolingual Sentence Embeddings L1 (English)

Parallel corpus L1–L2, with sentences s_i^{L1}, t_i^{L2} (or more languages)

Multilingual Sentence Embeddings with SBERT and MKD

Multilingual Knowledge Distillation

Idea

(Good) Monolingual Sentence Embeddings L1 (English) ⇒ **Teacher Model**

Parallel corpus L1–L2, with sentences s_i^{L1}, t_i^{L2} (or more languages)

Multilingual Sentence Embeddings with SBERT and MKD

Multilingual Knowledge Distillation

Idea

(Good) Monolingual Sentence Embeddings L1 (English) $\Rightarrow$ **Teacher Model**

Parallel corpus L1–L2, with sentences s_i^{L1}, t_i^{L2} (or more languages)

What do we want? $\mathrm{Embedding}(s_k^{L1}) \approx \mathrm{Embedding}(t_k^{L2})$

Multilingual Sentence Embeddings with SBERT and MKD

Multilingual Knowledge Distillation

Idea

(Good) Monolingual Sentence Embeddings L1 (English) $\Rightarrow$ **Teacher Model**

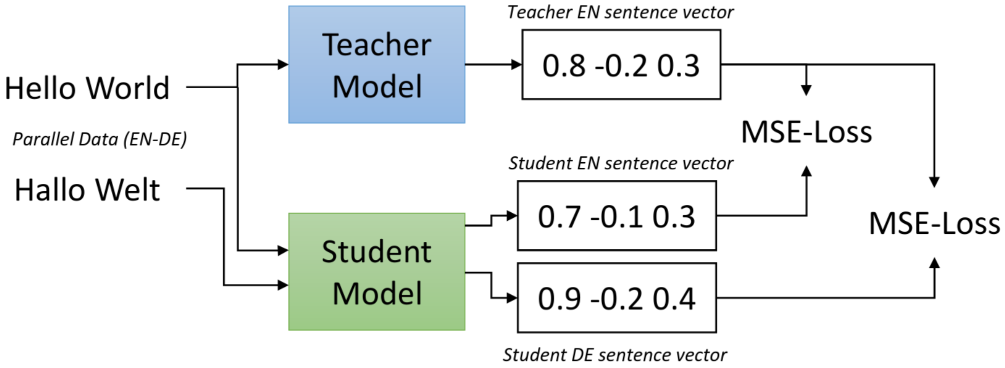
Parallel corpus L1–L2, with sentences s_i^{L1}, t_i^{L2} (or more languages)

What do we want? $\mathrm{Embedding}(s_k^{L1}) \approx \mathrm{Embedding}(t_k^{L2})$ $\Leftarrow$ **Student Model**

$M_{student}(s_k) \approx M_{teacher}(s_k)$ and $M_{student}(t_k) \approx M_{teacher}(s_k)$

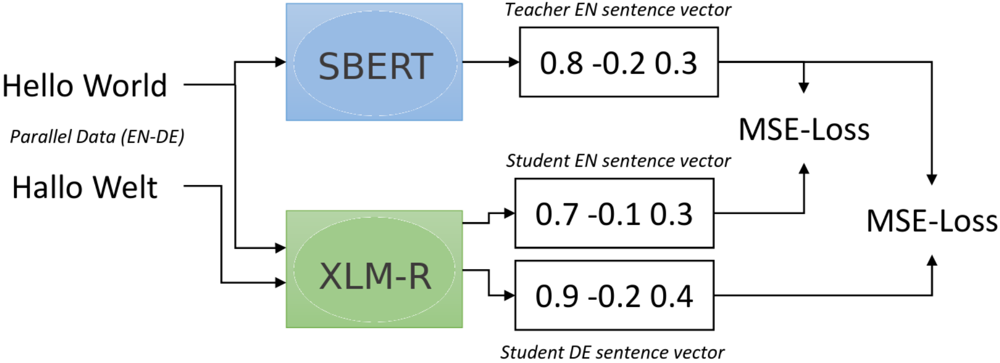
Multilingual Sentence Embeddings with SBERT and MKD

The Model



Multilingual Sentence Embeddings with SBERT and MKD

The Model



Multilingual Sentence Embeddings with SBERT and MKD

Multilingual Knowledge Distillation

Observations

$$L = \sum_k \left[(M_{student}(s_k) - M_{teacher}(s_k))^2 + (M_{student}(t_k) - M_{teacher}(s_k))^2 \right]$$

- vector space properties in the original source language from the teacher model are adopted and transferred to other languages
- vector spaces are aligned across languages, i.e., identical sentences in different languages are close

Multilingual Sentence Embeddings with SBERT and MKD

Multilingual Knowledge Distillation

Observations

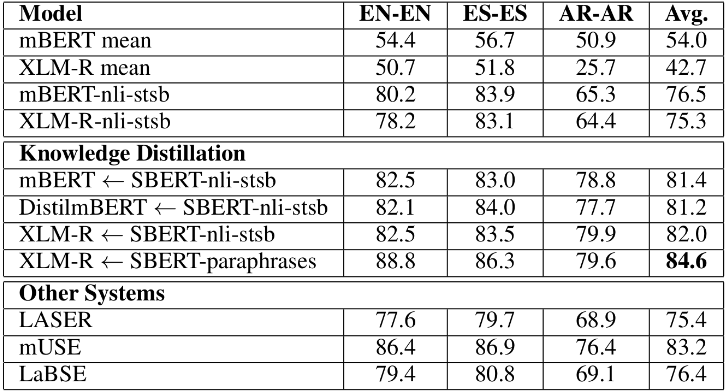
$$L = \sum_k \left[(M_{student}(s_k) - M_{teacher}(s_k))^2 + (M_{student}(t_k) - M_{teacher}(s_k))^2 \right]$$

- vector space properties in the original source language from the teacher model are adopted and transferred to other languages
- vector spaces are aligned across languages, i.e., identical sentences in different languages are close

- This is not necessary true for mBERT and XLM-RoBERTa (but they don't use parallel data)

Multilingual Sentence Embeddings with SBERT and MKD

MKD on STS Monolingual Pairs

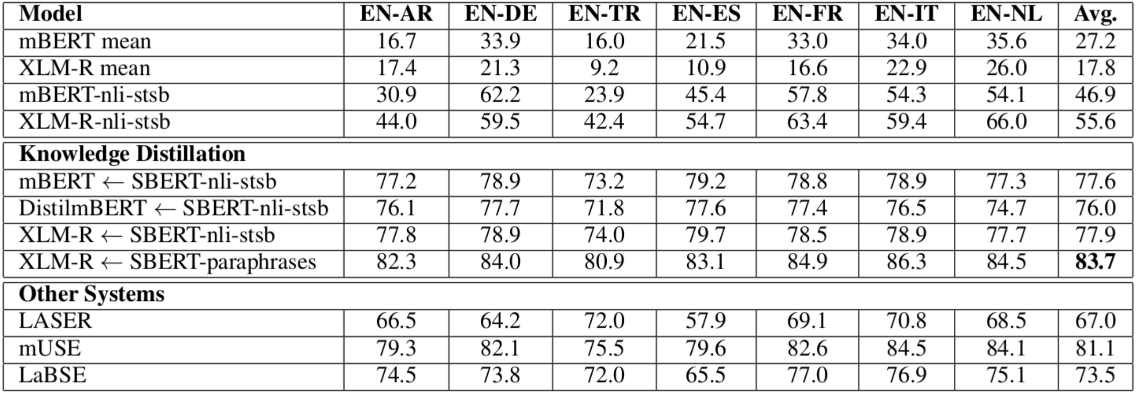


Model	**EN-EN**	**ES-ES**	**AR-AR**	**Avg.**
mBERT mean	54.4	56.7	50.9	54.0
XLM-R mean	50.7	51.8	25.7	42.7
mBERT-nli-stsb	80.2	83.9	65.3	76.5
XLM-R-nli-stsb	78.2	83.1	64.4	75.3
Knowledge Distillation				
mBERT ← SBERT-nli-stsb	82.5	83.0	78.8	81.4
DistilmBERT ← SBERT-nli-stsb	82.1	84.0	77.7	81.2
XLM-R ← SBERT-nli-stsb	82.5	83.5	79.9	82.0
XLM-R ← SBERT-paraphrases	88.8	86.3	79.6	**84.6**
Other Systems				
LASER	77.6	79.7	68.9	75.4
mUSE	86.4	86.9	76.4	83.2
LaBSE	79.4	80.8	69.1	76.4

- MKD improves base models, the true drop of mBERT and XML-R comes...

Multilingual Sentence Embeddings with SBERT and MKD

MKD on STS Cross-lingual Pairs

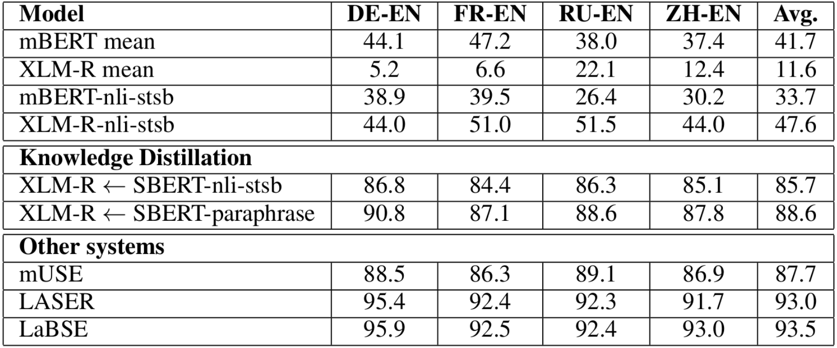


Model	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
mBERT mean	16.7	33.9	16.0	21.5	33.0	34.0	35.6	27.2
XLM-R mean	17.4	21.3	9.2	10.9	16.6	22.9	26.0	17.8
mBERT-nli-stsb	30.9	62.2	23.9	45.4	57.8	54.3	54.1	46.9
XLM-R-nli-stsb	44.0	59.5	42.4	54.7	63.4	59.4	66.0	55.6
Knowledge Distillation								
mBERT ← SBERT-nli-stsb	77.2	78.9	73.2	79.2	78.8	78.9	77.3	77.6
DistilmBERT ← SBERT-nli-stsb	76.1	77.7	71.8	77.6	77.4	76.5	74.7	76.0
XLM-R ← SBERT-nli-stsb	77.8	78.9	74.0	79.7	78.5	78.9	77.7	77.9
XLM-R ← SBERT-paraphrases	82.3	84.0	80.9	83.1	84.9	86.3	84.5	**83.7**
Other Systems								
LASER	66.5	64.2	72.0	57.9	69.1	70.8	68.5	67.0
mUSE	79.3	82.1	75.5	79.6	82.6	84.5	84.1	81.1
LaBSE	74.5	73.8	72.0	65.5	77.0	76.9	75.1	73.5

- In both settings LASER and family underperform (MT task for training)

Multilingual Sentence Embeddings with SBERT and MKD

MKD on Bitext Mining (BUCC)



Model	**DE-EN**	**FR-EN**	**RU-EN**	**ZH-EN**	**Avg.**
mBERT mean	44.1	47.2	38.0	37.4	41.7
XLM-R mean	5.2	6.6	22.1	12.4	11.6
mBERT-nli-stsb	38.9	39.5	26.4	30.2	33.7
XLM-R-nli-stsb	44.0	51.0	51.5	44.0	47.6
Knowledge Distillation					
XLM-R ← SBERT-nli-stsb	86.8	84.4	86.3	85.1	85.7
XLM-R ← SBERT-paraphrase	90.8	87.1	88.6	87.8	88.6
Other systems					
mUSE	88.5	86.3	89.1	86.9	87.7
LASER	95.4	92.4	92.3	91.7	93.0
LaBSE	95.9	92.5	92.4	93.0	93.5

Table 3: F_1 score on the BUCC bitext mining task.

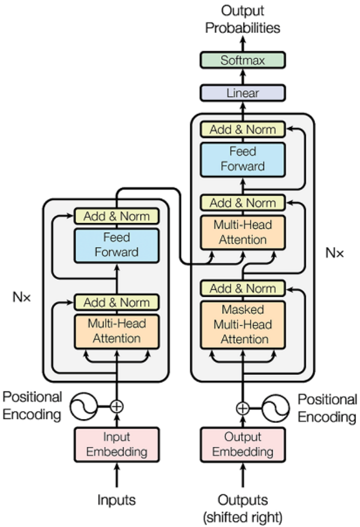
- LASER and family (MT task for training) outperform here

How are you doing? Need a Break?

Already a Long Way! And lots of Tables...

Multilingual Sentence Embeddings in NMT

Neural Machine Translation



(Vaswani et al., 2017)

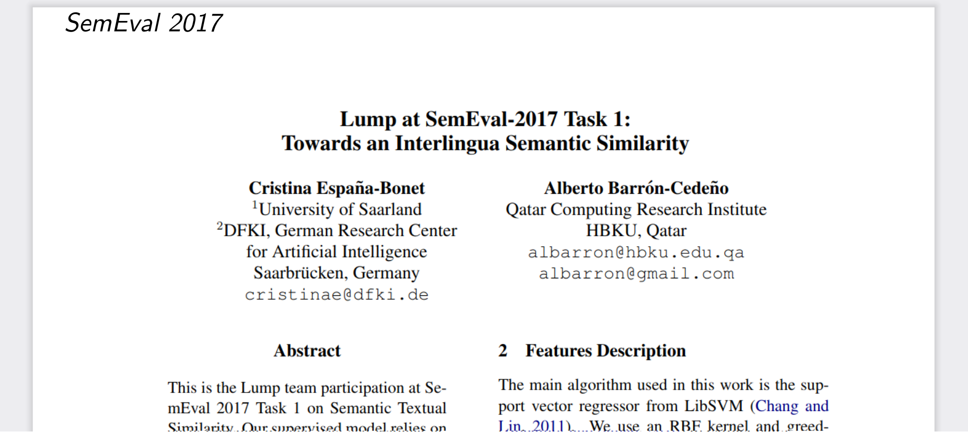
Multilingual Sentence Embeddings in NMT

Multilingual Neural Machine Translation, ML-NMT

- Machine translation is at least a bilingual task
- Neural machine translation encodes semantics in vectors
- Straightforward extension of NMT to multilingual NMT (ML-NMT)
- Simple architecture for ML-NMT: shared encoder & shared decoder
- ML word (or context) vectors lie in the same space

Multilingual Sentence Embeddings in NMT

Interlingua Semantic Similarity



SemEval 2017

Lump at SemEval-2017 Task 1: Towards an Interlingua Semantic Similarity

Cristina España-Bonet
[1]University of Saarland
[2]DFKI, German Research Center
for Artificial Intelligence
Saarbrücken, Germany
cristinae@dfki.de

Alberto Barrón-Cedeño
Qatar Computing Research Institute
HBKU, Qatar
albarron@hbku.edu.qa
albarron@gmail.com

Abstract

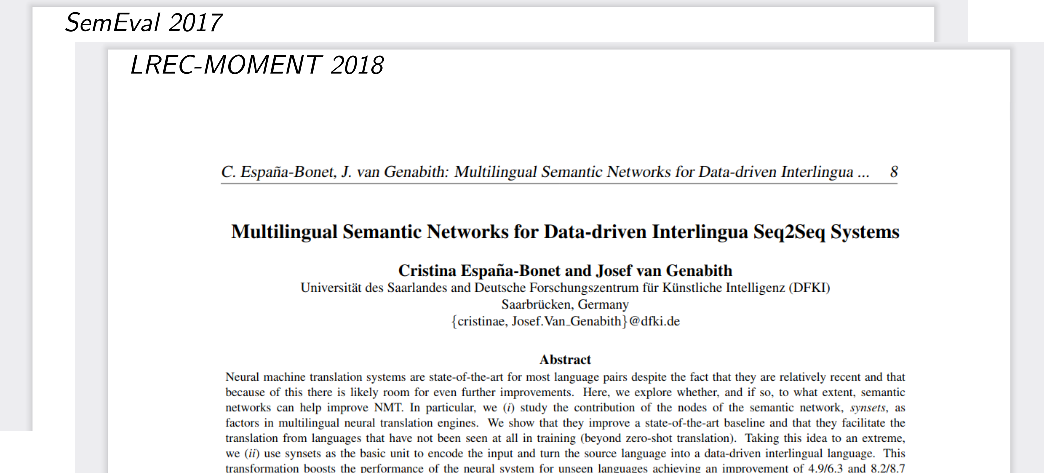
This is the Lump team participation at SemEval 2017 Task 1 on Semantic Textual Similarity. Our supervised model relies on

2 Features Description

The main algorithm used in this work is the support vector regressor from LibSVM (Chang and Lin, 2011). We use an RBF kernel and greed-

Multilingual Sentence Embeddings in NMT

Interlingua Semantic Similarity



SemEval 2017

LREC-MOMENT 2018

C. España-Bonet, J. van Genabith: Multilingual Semantic Networks for Data-driven Interlingua ... 8

Multilingual Semantic Networks for Data-driven Interlingua Seq2Seq Systems

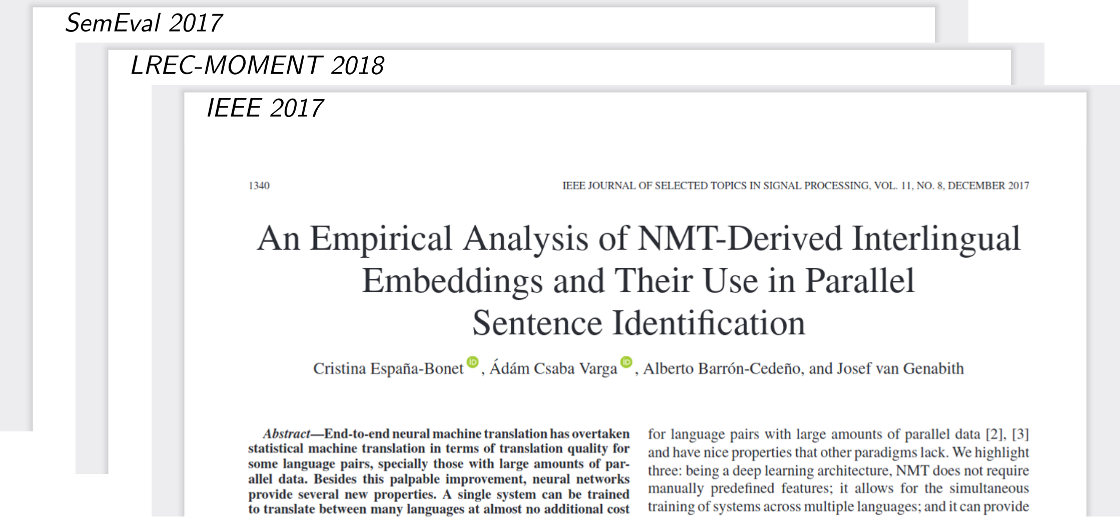
Cristina España-Bonet and Josef van Genabith
Universität des Saarlandes and Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI)
Saarbrücken, Germany
{cristinae, Josef.Van_Genabith}@dfki.de

Abstract

Neural machine translation systems are state-of-the-art for most language pairs despite the fact that they are relatively recent and that because of this there is likely room for even further improvements. Here, we explore whether, and if so, to what extent, semantic networks can help improve NMT. In particular, we (*i*) study the contribution of the nodes of the semantic network, *synsets*, as factors in multilingual neural translation engines. We show that they improve a state-of-the-art baseline and that they facilitate the translation from languages that have not been seen at all in training (beyond zero-shot translation). Taking this idea to an extreme, we (*ii*) use synsets as the basic unit to encode the input and turn the source language into a data-driven interlingual language. This transformation boosts the performance of the neural system for unseen languages achieving an improvement of 4.9/6.3 and 8.2/8.7

Multilingual Sentence Embeddings in NMT

Interlingua Semantic Similarity



SemEval 2017

LREC-MOMENT 2018

IEEE 2017

An Empirical Analysis of NMT-Derived Interlingual Embeddings and Their Use in Parallel Sentence Identification

Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith

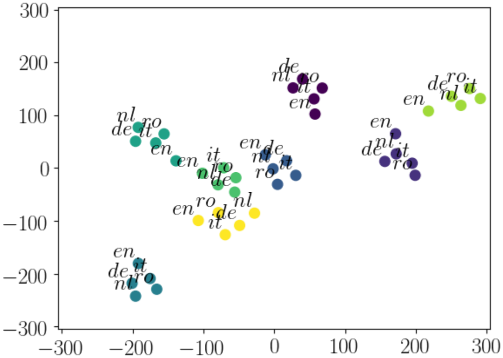
***Abstract*—End-to-end neural machine translation has overtaken statistical machine translation in terms of translation quality for some language pairs, specially those with large amounts of parallel data. Besides this palpable improvement, neural networks provide several new properties. A single system can be trained to translate between many languages at almost no additional cost**

for language pairs with large amounts of parallel data [2], [3] and have nice properties that other paradigms lack. We highlight three: being a deep learning architecture, NMT does not require manually predefined features; it allows for the simultaneous training of systems across multiple languages; and it can provide

Multilingual Sentence Embeddings in NMT

Multilingual Semantic Space for Context Vectors (easy)

(España-Bonet & van Genabith, 2018)

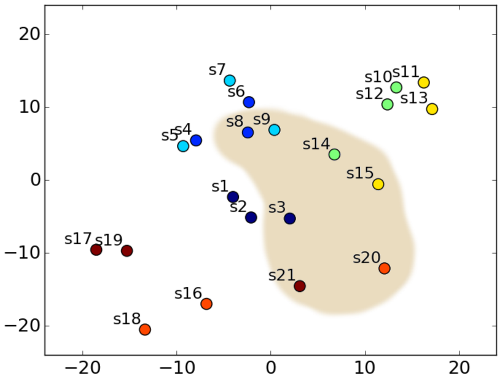


ML-NMT {*de*, *en*, *nl*, *it*, *ro*} → {*de*, *en*, *nl*, *it*, *ro*} with TED talks

Multilingual Sentence Embeddings in NMT

Multilingual Semantic Space for Context Vectors (hard)

(España-Bonet et al., 2017)



ML-NMT $\{en, es, ar\} \rightarrow \{en, es, ar\}$ with heterogeneous corpora

Multilingual Sentence Embeddings in NMT

Multilingual Semantic Space for Context Vectors (hard)

$s1$:$t1$ Spain princess testifies in historic fraud probe

$s2$:$t1$ Princesa de España testifica en juicio histórico de fraude

$s3$:$t1$ أميرة أسبانيا تدلي بشهادتها في قضية احتيال تاريخي.

$s4$:$t2$ You do not need to worry.

$s5$:$t3$ You don’t have to worry.

$s6$:$t2$ No necesitas preocuparte.

$s7$:$t3$ No te tienes por que preocupar.

$s8$:$t2$ لا ينبغي أن تقلق

$s9$:$t3$ لا ينبغي أن تجزع.

$s10$:$t4$ Mandela’s condition has ’improved’

$s11$:$t5$ Mandela’s condition has ’worsened over past 48 hours’

$s12$:$t4$ La salud de Mandela ha ’mejorado’

$s13$:$t5$ La salud de Mandela ’ha empeorado en las últimas 48 horas’

$s14$:$t4$ لقد تحسّنت حالة مانديلا الصحية.

$s15$:$t5$ ساءت الحالة الصحية لمانديلا خلال ال ٨٤ ساعة الماضية.

$s16$:$t6$ Vector space representation results in the loss of the order which the terms are in the document.

$s17$:$t7$ If a term occurs in the document, the value will be non-zero in the vector.

$s18$:$t6$ La representación en el espacio de vecores implica la pérdida del órden en el que los términos ocurren en el documento.

$s19$:$t7$ Si un término ocurre en el document, el valor en el vector será distinto de cero.

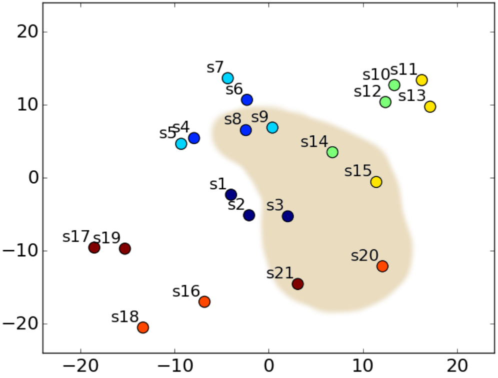
$s20$:$t6$ يؤدي تمثيلُ فضاءِ المتجهِ إلى فقد الترتيب الذي تكون عليه المصطلحات في الوثيقة.

$s21$:$t7$ إذا ما ورد مصطلح في الوثيقة، فالقيمة ستكون غيرصفرية المتجه.

Multilingual Sentence Embeddings in NMT

Multilingual Semantic Space for Context Vectors (hard)

(España-Bonet et al., 2017)

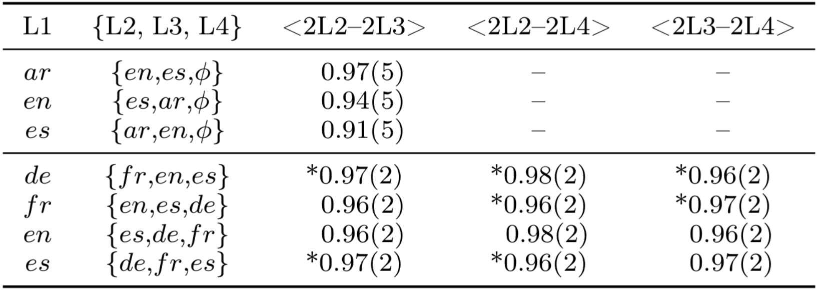


ML-NMT $\{en, es, ar\} \rightarrow \{en, es, ar\}$ with heterogeneous corpora

Multilingual Sentence Embeddings in NMT

How Close are Sentences Together?

Cosine similarities between the internal representations of the sentences in STS2017 and newstest2013 when translated from L1 into different languages L2, L3, L4.



L1	{L2, L3, L4}	<2L2–2L3>	<2L2–2L4>	<2L3–2L4>
ar	$\{en,es,\phi\}$	0.97(5)	–	–
en	$\{es,ar,\phi\}$	0.94(5)	–	–
es	$\{ar,en,\phi\}$	0.91(5)	–	–
de	$\{fr,en,es\}$	*0.97(2)	*0.98(2)	*0.96(2)
fr	$\{en,es,de\}$	0.96(2)	*0.96(2)	*0.97(2)
en	$\{es,de,fr\}$	0.96(2)	0.98(2)	0.96(2)
es	$\{de,fr,es\}$	*0.97(2)	*0.96(2)	0.97(2)

Multilingual Sentence Embeddings in NMT

Multilingual Semantic Space for Context Vectors

- Related languages cluster better together
  (for distant languages there might not even exist a mapping)

- The nature of the corpus also affects the clustering
  (corpus in different domains per language make the learning more difficult)

- These trends are common in several NLP tasks

Multilingual Sentence Embeddings in NMT

Multilingual Semantic Space for Context Vectors

- Related languages cluster better together
  (for distant languages there might not even exist a mapping)

- The nature of the corpus also affects the clustering
  (corpus in different domains per language make the learning more difficult)

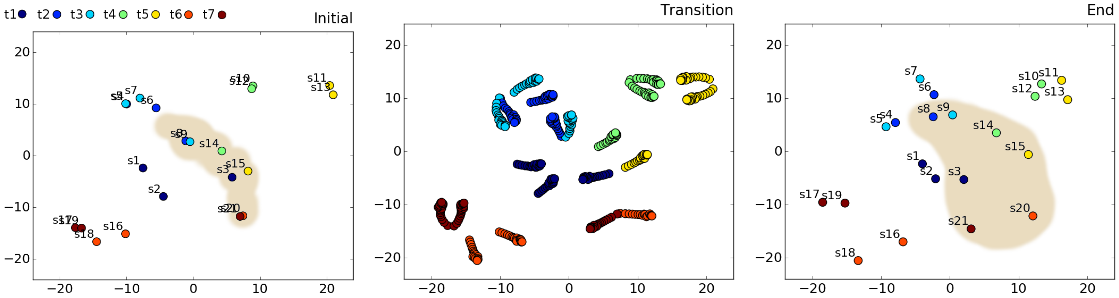
- These trends are common in several NLP tasks

- **What happens during training?**

Multilingual Sentence Embeddings in NMT

Evolution of Context Vectors through Training (hard)

(España-Bonet et al., 2017)



ML-NMT $\{en, es, ar\} \rightarrow \{en, es, ar\}$ with heterogeneous corpora

Multilingual Sentence Embeddings in NMT

Evolution of Context Vectors through Training (hard)

Pearson correlation on STS 2017 data

	track1 *ar–ar*	track2 *ar–en*	track3 *es–es*	track4a *es–en*	track5 *en–en*
`WE-d300`	0.49	0.28	0.55	0.40	0.56
`WE-d1024`	0.51	0.33	0.59	0.45	0.60

(España-Bonet & Barrón-Cedeño, 2017)

Multilingual Sentence Embeddings in NMT

Evolution of Context Vectors through Training (hard)

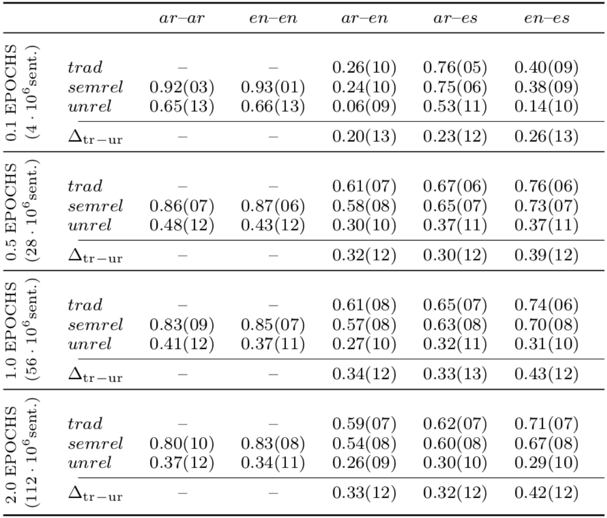
Pearson correlation on STS 2017 data

	track1 *ar–ar*	track2 *ar–en*	track3 *es–es*	track4a *es–en*	track5 *en–en*
`WE-d300`	0.49	0.28	0.55	0.40	0.56
`WE-d1024`	0.51	0.33	0.59	0.45	0.60
$\texttt{NMT}_{\text{ctx}}$-0.1Ep	0.32	0.25	0.55	0.32	0.54
$\texttt{NMT}_{\text{ctx}}$-0.5Ep	0.52	0.36	0.71	0.40	0.68
$\texttt{NMT}_{\text{ctx}}$-1.0Ep	0.57	0.42	0.74	0.44	0.72
$\texttt{NMT}_{\text{ctx}}$-2.0Ep	0.59	0.44	0.78	0.49	0.76

(España-Bonet & Barrón-Cedeño, 2017)

Multilingual Sentence Embeddings in NMT

Evolution According to the Similarity: from Translations to Unrelated Sentences



		ar–ar	*en–en*	*ar–en*	*ar–es*	*en–es*
0.1 EPOCHS ($4 \cdot 10^6$sent.)	*trad*	–	–	0.26(10)	0.76(05)	0.40(09)
	semrel	0.92(03)	0.93(01)	0.24(10)	0.75(06)	0.38(09)
	unrel	0.65(13)	0.66(13)	0.06(09)	0.53(11)	0.14(10)
	$\Delta_{\mathrm{tr-ur}}$	–	–	0.20(13)	0.23(12)	0.26(13)
0.5 EPOCHS ($28 \cdot 10^6$sent.)	*trad*	–	–	0.61(07)	0.67(06)	0.76(06)
	semrel	0.86(07)	0.87(06)	0.58(08)	0.65(07)	0.73(07)
	unrel	0.48(12)	0.43(12)	0.30(10)	0.37(11)	0.37(11)
	$\Delta_{\mathrm{tr-ur}}$	–	–	0.32(12)	0.30(12)	0.39(12)
1.0 EPOCHS ($56 \cdot 10^6$sent.)	*trad*	–	–	0.61(08)	0.65(07)	0.74(06)
	semrel	0.83(09)	0.85(07)	0.57(08)	0.63(08)	0.70(08)
	unrel	0.41(12)	0.37(11)	0.27(10)	0.32(11)	0.31(10)
	$\Delta_{\mathrm{tr-ur}}$	–	–	0.34(12)	0.33(13)	0.43(12)
2.0 EPOCHS ($112 \cdot 10^6$sent.)	*trad*	–	–	0.59(07)	0.62(07)	0.71(07)
	semrel	0.80(10)	0.83(08)	0.54(08)	0.60(08)	0.67(08)
	unrel	0.37(12)	0.34(11)	0.26(09)	0.30(10)	0.29(10)
	$\Delta_{\mathrm{tr-ur}}$	–	–	0.33(12)	0.32(12)	0.42(12)

Cosine similarities between the obtained representations of the sentences in the STS2017 test set

trad: sim 5
semrel: sim 4
unrel: sim 0

Multilingual Sentence Embeddings in NMT

Semantic Language-independent Clustering in ML-NMT

This is a fact. ML-NMT behaves this way.

Can we profit from it?

Outline

1. Motivation

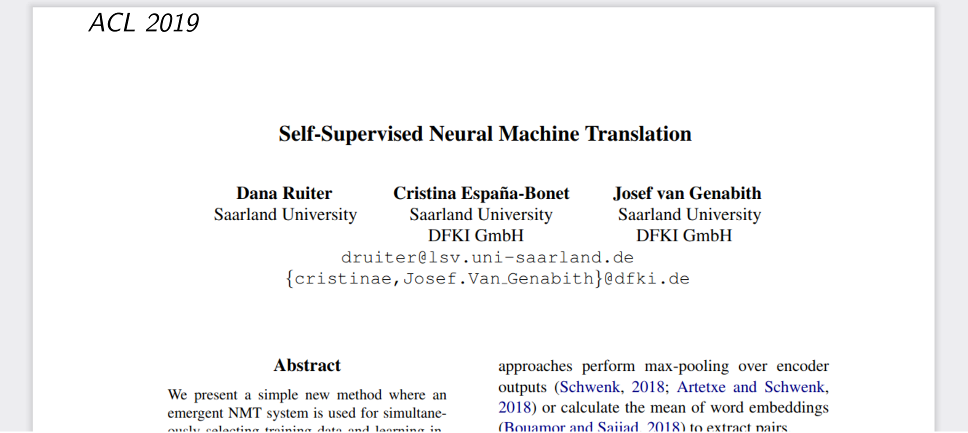
2. (Multilingual) Sentence Embeddings

3. Self-Supervised NMT

4. Initialising (Multilingual) NMT

Self-Supervised NMT

Exploiting the Evolution of NMT Embeddings



ACL 2019

Self-Supervised Neural Machine Translation

Dana Ruiter
Saarland University

Cristina España-Bonet
Saarland University
DFKI GmbH

Josef van Genabith
Saarland University
DFKI GmbH

druiter@lsv.uni-saarland.de
{cristinae,Josef.Van_Genabith}@dfki.de

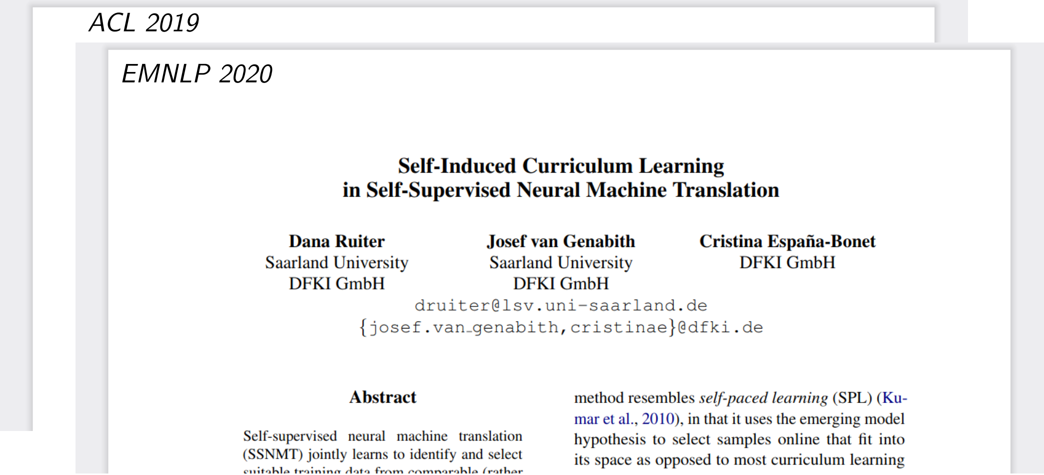
Abstract

We present a simple new method where an emergent NMT system is used for simultaneously selecting training data and learning in

approaches perform max-pooling over encoder outputs (Schwenk, 2018; Artetxe and Schwenk, 2018) or calculate the mean of word embeddings (Bouamor and Sajjad, 2018) to extract pairs

Self-Supervised NMT

Exploiting the Evolution of NMT Embeddings



ACL 2019

EMNLP 2020

Self-Induced Curriculum Learning in Self-Supervised Neural Machine Translation

Dana Ruiter
Saarland University
DFKI GmbH

Josef van Genabith
Saarland University
DFKI GmbH

Cristina España-Bonet
DFKI GmbH

druiter@lsv.uni-saarland.de
{josef.van_genabith,cristinae}@dfki.de

Abstract

Self-supervised neural machine translation (SSNMT) jointly learns to identify and select suitable training data from comparable (rather

method resembles *self-paced learning* (SPL) (Kumar et al., 2010), in that it uses the emerging model hypothesis to select samples online that fit into its space as opposed to most curriculum learning

Self-Supervised NMT

Question

- NMT training differentiates translations from non-translations very soon
- In a standard NMT, all training sentences are (should be) translations
- Can we feed the system with any kind of sentence pair and let itself decide if it is useful or not?

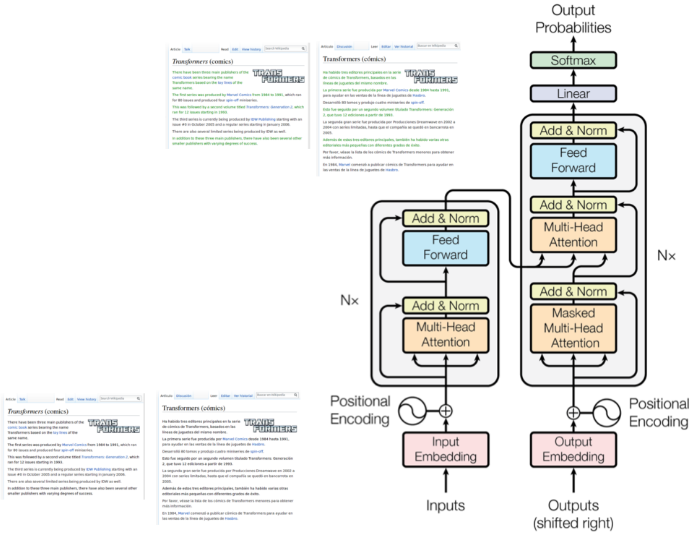
Self-Supervised NMT

Question

- NMT training differentiates translations from non-translations very soon
- In a standard NMT, all training sentences are (should be) translations
- Can we feed the system with any kind of sentence pair and let itself decide if it is useful or not?
- **Yes, we can!**

Self-Supervised NMT

Main Idea



Article Talk Read Edit View history Search Wikipedia

Transformers (comics)

There have been three main publishers of the comic book series bearing the name Transformers based on the toy lines of the same name.

The first series was produced by Marvel Comics from 1984 to 1991, which ran for 80 issues and produced four spin-off miniseries.

This was followed by a second volume titled *Transformers: Generation 2*, which ran for 12 issues starting in 1993.

The third series is currently being produced by IDW Publishing starting with an issue #0 in October 2005 and a regular series starting in January 2006.

There are also several limited series being produced by IDW as well.

In addition to these three main publishers, there have also been several other smaller publishers with varying degrees of success.

Artículo Discusión Leer Editar Ver historial Buscar en Wikipedia

Transformers (cómics)

Ha habido tres editores principales en la serie de cómics de Transformers, basados en las líneas de juguetes del mismo nombre.

La primera serie fue producida por Marvel Comics desde 1984 hasta 1991, para ayudar en las ventas de la línea de juguetes de Hasbro.

Desarrolló 80 tomos y produjo cuatro miniseries de spin-off.

Esto fue seguido por un segundo volumen titulado Transformers: Generación 2, que tuvo 12 ediciones a partir de 1993.

La segunda gran serie fue producida por Producciones Dreamwave en 2002 a 2004 con series limitadas, hasta que el compañía se quedó en bancarrota en 2005.

Además de estos tres editores principales, también ha habido varias otras editoriales más pequeñas con diferentes grados de éxito.

Por favor, véase la lista de los cómics de Transformers menores para obtener más información.

En 1984, Marvel comenzó a publicar cómics de Transformers para ayudar en las ventas de la línea de juguetes de Hasbro.

Article Talk Read Edit View history Search Wikipedia

Transformers (comics)

There have been three main publishers of the comic book series bearing the name Transformers based on the toy lines of the same name.

The first series was produced by Marvel Comics from 1984 to 1991, which ran for 80 issues and produced four spin-off miniseries.

This was followed by a second volume titled *Transformers: Generation 2*, which ran for 12 issues starting in 1993.

The third series is currently being produced by IDW Publishing starting with an issue #0 in October 2005 and a regular series starting in January 2006.

There are also several limited series being produced by IDW as well.

In addition to these three main publishers, there have also been several other smaller publishers with varying degrees of success.

Artículo Discusión Leer Editar Ver historial Buscar en Wikipedia

Transformers (cómics)

Ha habido tres editores principales en la serie de cómics de Transformers, basados en las líneas de juguetes del mismo nombre.

La primera serie fue producida por Marvel Comics desde 1984 hasta 1991, para ayudar en las ventas de la línea de juguetes de Hasbro.

Desarrolló 80 tomos y produjo cuatro miniseries de spin-off.

Esto fue seguido por un segundo volumen titulado Transformers: Generación 2, que tuvo 12 ediciones a partir de 1993.

La segunda gran serie fue producida por Producciones Dreamwave en 2002 a 2004 con series limitadas, hasta que el compañía se quedó en bancarrota en 2005.

Además de estos tres editores principales, también ha habido varias otras editoriales más pequeñas con diferentes grados de éxito.

Por favor, véase la lista de los cómics de Transformers menores para obtener más información.

En 1984, Marvel comenzó a publicar cómics de Transformers para ayudar en las ventas de la línea de juguetes de Hasbro.

Self-Supervised NMT

Main Idea

- Parallel data extraction as an auxiliary task to enable NMT training
- NMT training as an auxiliary task to enhance parallel sentence extraction

Self-Supervised NMT

Main Idea

- Parallel data extraction as an auxiliary task to enable NMT training
- NMT training as an auxiliary task to enhance parallel sentence extraction

Self-supervision?

Just in a non-standard way, none of the tasks is completely supervised

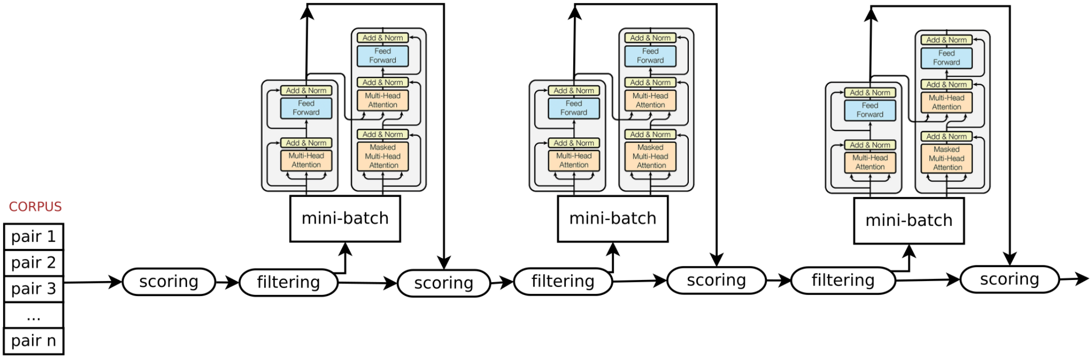
Self-Supervised NMT

Main Idea

- Joint selection of sentences & training NMT
- Uses internal embeddings, i.e., architecture independent
- Bidirectional training {L1, L2}→{L1, L2} (shared encoder)
- Optional initialisation with word embeddings trained on monolingual corpora
- On-line process: embeddings change through epochs, therefore selected sentences change through epochs

Self-Supervised NMT

Training Procedure



CORPUS

pair 1

pair 2

pair 3

...

pair n

scoring → filtering → scoring → filtering → scoring → filtering → scoring →

mini-batch

mini-batch

mini-batch

Self-Supervised NMT

Algorithm Description

1. Internal NMT representation: E_w (words); E_h (sentence)
2. Score all sentence pairs in a lot (i.e. WP article)
3. Filter options
4. Add filtered sentences into a mini-batch
5. Train system when mini-batch is complete
6. Update weights and continue with more data and go again to 1.

Self-Supervised NMT

Joint Training: Key Points

1. **Sentence Representation**

2. **Scoring function**

Self-Supervised NMT

Joint Training: Key Points

1. **Sentence Representation**
   the sum of word embeddings (E_w) and the hidden states in an RNN or the encoder outputs in a transformer (E_h):

$$E_w = \sum_{t=1}^{T} e_t, \qquad E_h = \sum_{t=1}^{T} h_t$$

2. **Scoring function**

Self-Supervised NMT

Joint Training: Key Points

1. **Sentence Representation**
   S_{L1} and S_{L2} vector representations for each sentence of a pair (E_w or E_h)

2. **Scoring function**

   cosine similarity: $$\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \frac{S_{\mathrm{L1}} \cdot S_{\mathrm{L2}}}{\|S_{\mathrm{L1}}\| \, \|S_{\mathrm{L2}}\|}$$

   margin-based score:

   $$\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}}) = \frac{\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})}{\mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L1}}, P_k)/2 + \mathrm{avr}_{\mathrm{kNN}}(S_{\mathrm{L2}}, Q_k)/2}$$

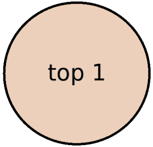
   where $$\mathrm{avr}_{\mathrm{kNN}}(X, Y_k) = \sum_{Y \in k\mathrm{NN}(X)} \frac{\cos(X, Y)}{k}$$ (average similarity)

Self-Supervised NMT

Sentence Selection (Filtering)

1. Input a lot (e.g. set of WP article pairs, web pages, etc)
2. Score all sentence pairs
3. Keep the top one pairs (with constraints!)

E_h src2tgt

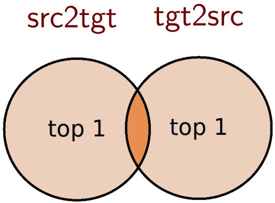


Self-Supervised NMT

Sentence Selection (Filtering)

1. Input a lot (e.g. set of WP article pairs, web pages, etc)
2. Score all sentence pairs
3. Keep the top one pairs (with constraints!)

E_h

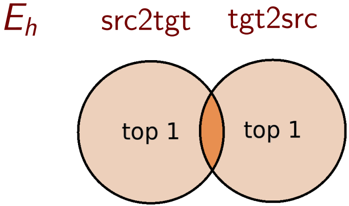


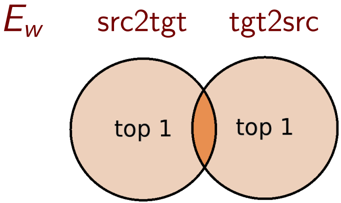
Self-Supervised NMT

Sentence Selection (Filtering)

1. Input a lot (e.g. set of WP article pairs, web pages, etc)
2. Score all sentence pairs

3. Keep the top one pairs (with constraints!)

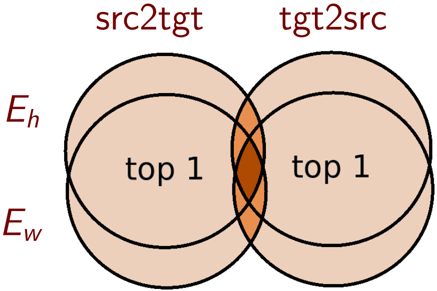




Self-Supervised NMT

Sentence Selection (Filtering)

Intersection of intersection of intersection...



to avoid the need for a threshold
(remember LASER bitext mining approach)

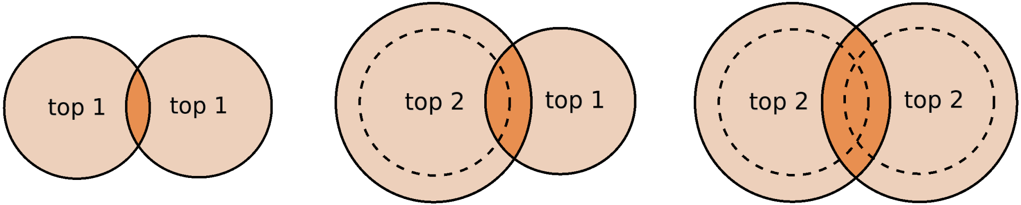
Self-Supervised NMT

Sentence Selection: Precision or Recall?

low permissibility

medium permissibility

high permissibility



high precision mode

high recall mode

Self-Supervised NMT

Models: Transformer Encoders

cosP: E_w, E_h in high precision mode and $\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ are used.

margP: E_w, E_h in high precision mode and $\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ are used.

Self-Supervised NMT

Models: Transformer Encoders

cosP: E_w, E_h in high precision mode and $\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ are used.

margP: E_w, E_h in high precision mode and $\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ are used.

margR: As **margP** but E_w and E_h are used in the high recall mode.

Self-Supervised NMT

Models: Transformer Encoders

cosP: E_w, E_h in high precision mode and $\cos(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ are used.

margP: E_w, E_h in high precision mode and $\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$ are used.

margR: As **margP** but E_w and E_h are used in the high recall mode.

margH: As **margP** with E_h as only representation.
A hard threshold of 1.01 is used.

margE: As **margP** with E_w as only representation.
A hard threshold of 1.00 is used.

SS-NMT: Detailed Results on *fr-en* with Wikipedia

Performance as Measured by BLEU

Model	Corpus, *en*+*fr* sent. (in millions)	BLEU *en2fr*	 *fr2en*
cosP	Wikipedia, 12+8	25.21	24.96
margE	Wikipedia, 12+8	27.33	25.87
margH	Wikipedia, 12+8	24.45	23.83
margP	Wikipedia, 12+8	**29.21**	**27.36**
margR	Wikipedia, 12+8	28.01	26.78

margP: E_w, E_h in high precision mode and $\mathrm{margin}(S_{\mathrm{L1}}, S_{\mathrm{L2}})$

SS-NMT: Automatic Evaluation

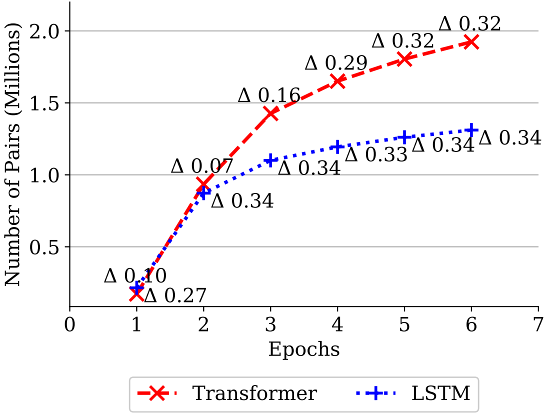
Comparison with Unsupervised NMT

	SS-NMT						SotA	
	L1-to-L2			L2-to-L1			L1-to-L2	L2-to-L1
L1–L2	BLEU	TER	METEOR	BLEU	TER	METEOR	BLEU	BLEU
en–fr	29.5±.6	51.9±.6	46.4±.6	27.7±.6	53.4±.7	30.3±.4	45.6/25.1/37.5	—-/24.2/34.9
en–de	15.2±.5	68.5±.7	30.3±.5	21.2±.6	62.8±.9	25.4±.4	37.9/17.2/28.3	—-/21.0/35.2
en–es	28.6±.7	52.6±.7	47.8±.7	28.4±.7	54.1±.7	30.5±.4	–/–/–	–/–/–

Scores on Newstest 2014 (*fr*) Newstest 2016 (*de*) and Newstest 2013 (*es*). Comparison with three SotA systems for supervised NMT (Edunov et al. 2018) / USNMT (Lample et al. 2018) / pre-trained+LM USNMT (Song et al. 2019)

SS-NMT: Behaviour through Training

What's going on? — margP *models*



- The mean difference in similarity between accepted and rejected pairs increases (Δ)
- The number of extracted sentences increases with Δ
- Changes are more prominent at the beginning of the training

SS-NMT: Behaviour through Training

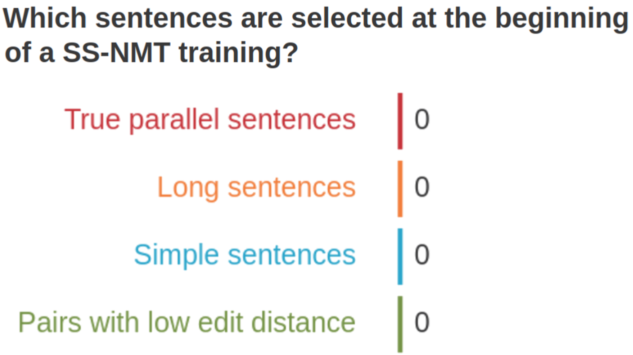
Built-In Curriculum

	#Pairs$_{enfr}$	*en2fr*	*fr2en*	#Pairs$_{ende}$	*en2de*	*de2en*	#Pairs$_{enes}$	*en2es*	*es2en*
NMT$_{init}$	2.14M	21.8±.6	21.1±.5	0.32M	3.4±.3	4.7±.3	2.51M	27.0±.7	25.0±.7
NMT$_{mid}$	3.14M	29.0±.6	26.6±.6	1.13M	11.2±.4	15.0±.6	3.96M	28.3±.7	26.1±.7
NMT$_{end}$	3.17M	28.8±.6	26.5±.6	1.18M	11.9±.5	15.3±.5	3.99M	28.3±.7	26.2±.7
NMT$_{all}$	5.38M	26.8±.7	25.2±.6	2.21M	11.6±.5	15.0±.6	5.41M	27.9±.6	25.9±.8
SS-NMT	5.38M	29.5±.6	27.7±.6	2.21M	14.4±.6	18.1±.6	5.41M	28.6±.7	28.4±.7

Supervised NMT systems trained on the unique pairs collected by SS-NMT in the first (NMT$_{init}$), intermediate (NMT$_{mid}$), final (NMT$_{end}$) and all (NMT$_{all}$) epochs of training

Learning Process in SS-NMT

What's your Intuition? (DirectPoll)



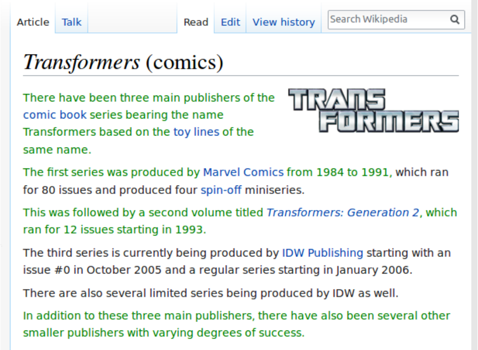
Which sentences are selected at the beginning of a SS-NMT training?

True parallel sentences	0
Long sentences	0
Simple sentences	0
Pairs with low edit distance	0

Learning Process in SS-NMT

What's going on? — Built-In Curriculum Learning

Input Documents



Transformers (comics)

There have been three main publishers of the comic book series bearing the name Transformers based on the toy lines of the same name.

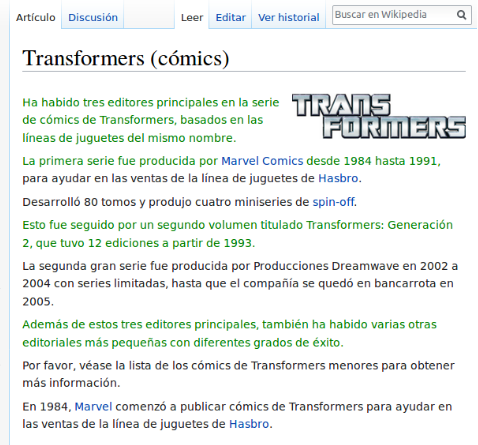
The first series was produced by Marvel Comics from 1984 to 1991, which ran for 80 issues and produced four spin-off miniseries.

This was followed by a second volume titled *Transformers: Generation 2*, which ran for 12 issues starting in 1993.

The third series is currently being produced by IDW Publishing starting with an issue #0 in October 2005 and a regular series starting in January 2006.

There are also several limited series being produced by IDW as well.

In addition to these three main publishers, there have also been several other smaller publishers with varying degrees of success.



Transformers (cómics)

Ha habido tres editores principales en la serie de cómics de Transformers, basados en las líneas de juguetes del mismo nombre.

La primera serie fue producida por Marvel Comics desde 1984 hasta 1991, para ayudar en las ventas de la línea de juguetes de Hasbro.

Desarrolló 80 tomos y produjo cuatro miniseries de spin-off.

Esto fue seguido por un segundo volumen titulado Transformers: Generación 2, que tuvo 12 ediciones a partir de 1993.

La segunda gran serie fue producida por Producciones Dreamwave en 2002 a 2004 con series limitadas, hasta que el compañía se quedó en bancarrota en 2005.

Además de estos tres editores principales, también ha habido varias otras editoriales más pequeñas con diferentes grados de éxito.

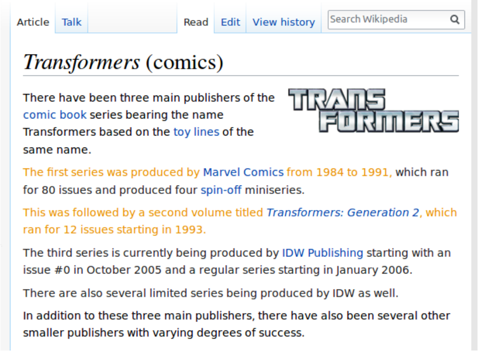
Por favor, véase la lista de los cómics de Transformers menores para obtener más información.

En 1984, Marvel comenzó a publicar cómics de Transformers para ayudar en las ventas de la línea de juguetes de Hasbro.

Learning Process in SS-NMT

Built-In Curriculum Learning

Sentence selection through epochs: Epoch 1



Transformers (comics)

There have been three main publishers of the comic book series bearing the name Transformers based on the toy lines of the same name.

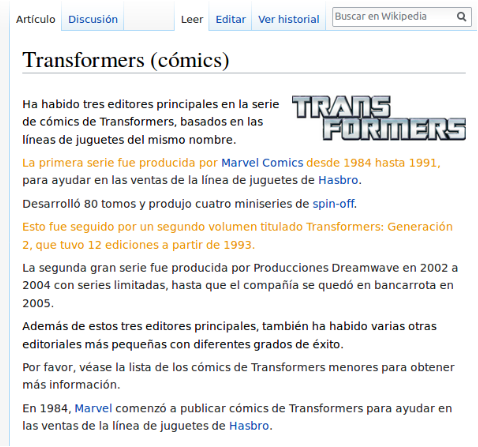
The first series was produced by Marvel Comics from 1984 to 1991, which ran for 80 issues and produced four spin-off miniseries.

This was followed by a second volume titled *Transformers: Generation 2*, which ran for 12 issues starting in 1993.

The third series is currently being produced by IDW Publishing starting with an issue #0 in October 2005 and a regular series starting in January 2006.

There are also several limited series being produced by IDW as well.

In addition to these three main publishers, there have also been several other smaller publishers with varying degrees of success.



Transformers (cómics)

Ha habido tres editores principales en la serie de cómics de Transformers, basados en las líneas de juguetes del mismo nombre.

La primera serie fue producida por Marvel Comics desde 1984 hasta 1991, para ayudar en las ventas de la línea de juguetes de Hasbro.

Desarrolló 80 tomos y produjo cuatro miniseries de spin-off.

Esto fue seguido por un segundo volumen titulado Transformers: Generación 2, que tuvo 12 ediciones a partir de 1993.

La segunda gran serie fue producida por Producciones Dreamwave en 2002 a 2004 con series limitadas, hasta que el compañía se quedó en bancarrota en 2005.

Además de estos tres editores principales, también ha habido varias otras editoriales más pequeñas con diferentes grados de éxito.

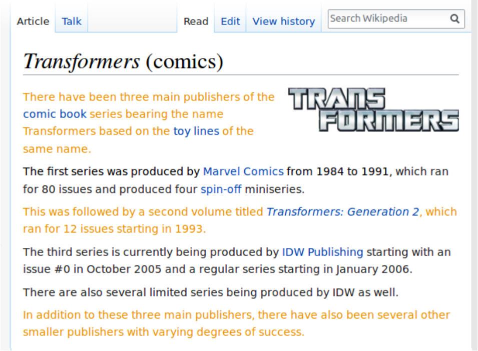
Por favor, véase la lista de los cómics de Transformers menores para obtener más información.

En 1984, Marvel comenzó a publicar cómics de Transformers para ayudar en las ventas de la línea de juguetes de Hasbro.

Learning Process in SS-NMT

Built-In Curriculum Learning

Sentence selection through epochs: Epoch 6



Transformers (comics)

There have been three main publishers of the comic book series bearing the name Transformers based on the toy lines of the same name.

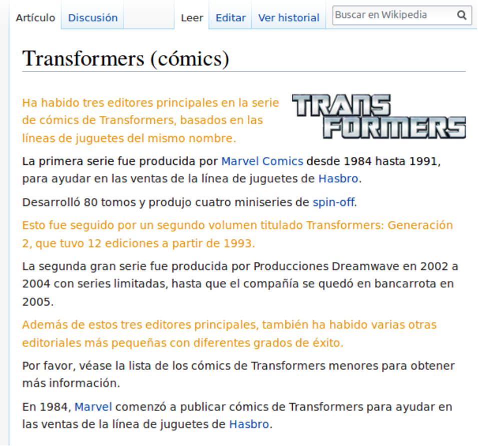
The first series was produced by Marvel Comics from 1984 to 1991, which ran for 80 issues and produced four spin-off miniseries.

This was followed by a second volume titled *Transformers: Generation 2*, which ran for 12 issues starting in 1993.

The third series is currently being produced by IDW Publishing starting with an issue #0 in October 2005 and a regular series starting in January 2006.

There are also several limited series being produced by IDW as well.

In addition to these three main publishers, there have also been several other smaller publishers with varying degrees of success.



Transformers (cómics)

Ha habido tres editores principales en la serie de cómics de Transformers, basados en las líneas de juguetes del mismo nombre.

La primera serie fue producida por Marvel Comics desde 1984 hasta 1991, para ayudar en las ventas de la línea de juguetes de Hasbro.

Desarrolló 80 tomos y produjo cuatro miniseries de spin-off.

Esto fue seguido por un segundo volumen titulado Transformers: Generación 2, que tuvo 12 ediciones a partir de 1993.

La segunda gran serie fue producida por Producciones Dreamwave en 2002 a 2004 con series limitadas, hasta que el compañía se quedó en bancarrota en 2005.

Además de estos tres editores principales, también ha habido varias otras editoriales más pequeñas con diferentes grados de éxito.

Por favor, véase la lista de los cómics de Transformers menores para obtener más información.

En 1984, Marvel comenzó a publicar cómics de Transformers para ayudar en las ventas de la línea de juguetes de Hasbro.

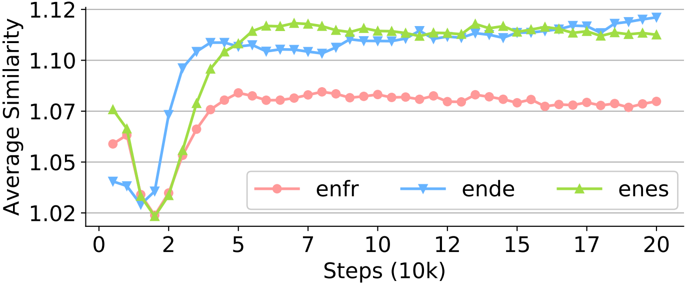
Learning Process in SS-NMT

Self-Induced Curricula

- SS-NMT induces a curriculum when selecting the data to train the MT task
- The order in which sentences are extracted is vital for translation quality (NMTall vs. SS-NMT)
- The data selection shows (at least) 3 curricula:
  1. a task-specific (MT) curriculum
  2. a denoising curriculum
  3. a complexity curriculum

Self-Induced Curricula

Task-specific (MT) Curriculum

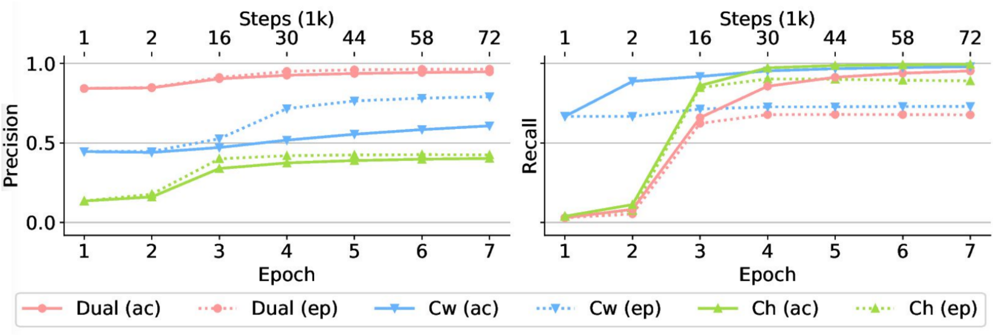


⇝ more cross-lingual similarity → more parallel

⇝ more parallel → closer to MT purpose

Self-Induced Curricula

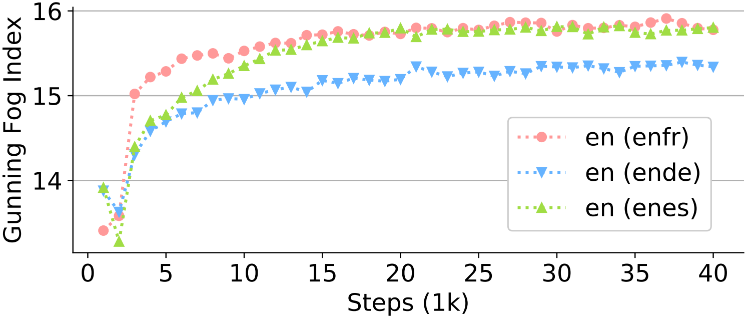
Denoising Curriculum



- Need of a synthetic corpus (scrambled Europarl)
- The percentage of non-matching pairs, i.e. non-translations, decreases from 18% to 2% (*en2fr*)

Self-Induced Curricula

Complexity Curriculum

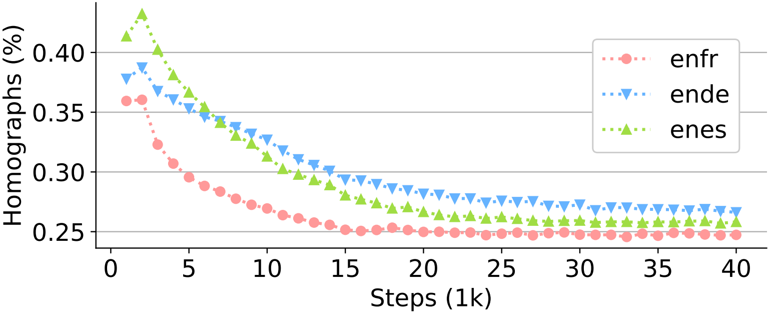


Gunning Fog, readability measure: $\mathrm{GF} = 0.4\left[\left(\frac{w}{s}\right) + 100\left(\frac{c}{w}\right)\right]$

- Increment from GF=11 (high school students) to GF=13 (undergrads)

Self-Induced Curricula

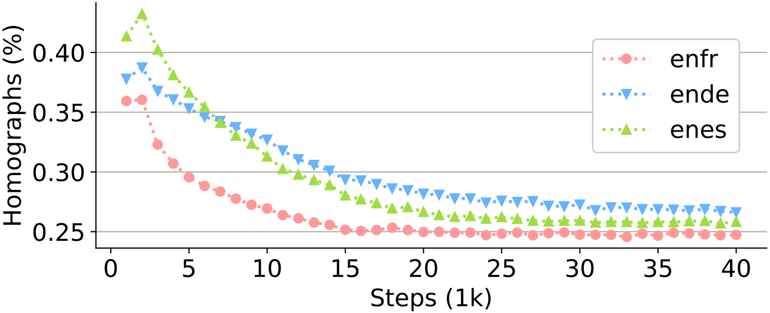
Key Point: Homographs!



- Large % of homographs in the sentences at the beginning of the training less sentences (punctuation, numbers, common BPE), noisier, easier

Self-Induced Curricula

Key Point: Homographs!



- Large % of homographs in the sentences at the beginning of the training less sentences (punctuation, numbers, common BPE), noisier, easier

⇝ What if no homographs?

Self-Induced Curricula

Open Problems

1. **Distant Languages** (no/few homographs)
2. **Low-resourced languages**

Similar issues in unsupervised NMT.

Same solutions?

SS-NMT: Low-resource Setting

On-going Work

On-line back-translation of rejected pairs:

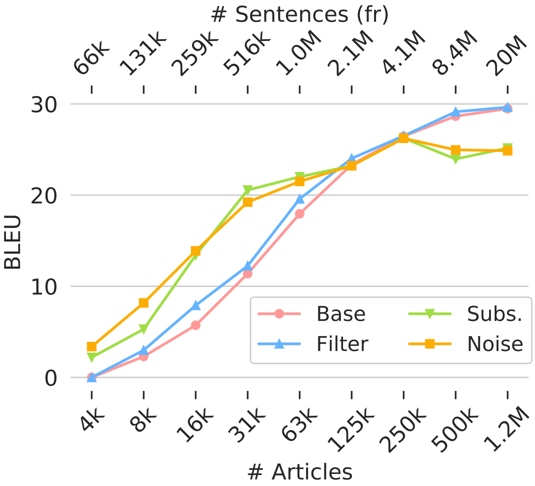
- SS-NMT filtering to remove low-quality back-translations
- Word translation for rejected back-translations
- Add noise (word removal, replacement and permutation)

Performance:

- Artificial setting 👍 (lots of mono data, few comparable)
- Real setting 👎 (few mono data, few comparable)

SS-NMT: Low-resource Setting

On-going Work



- Damages high-resource setting
- Significant improvements mid-resource setting
- Small improvements in the low-resource setting

Outline

1. Motivation

2. (Multilingual) Sentence Embeddings

3. Self-Supervised NMT

4. Initialising (Multilingual) NMT

Initialising (Multilingual) NMT

Remember... NMT with Transformers:

embeddings ⇝

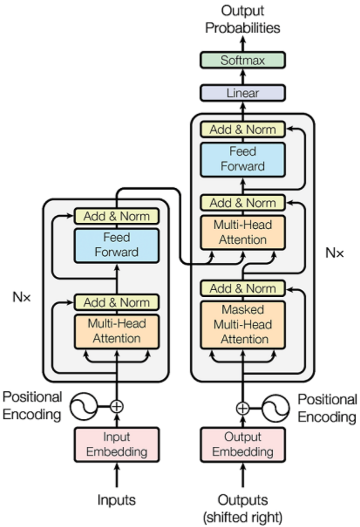
embeddings ⇝

embeddings ⇝

embeddings ⇝

embeddings ⇝

embeddings ⇝



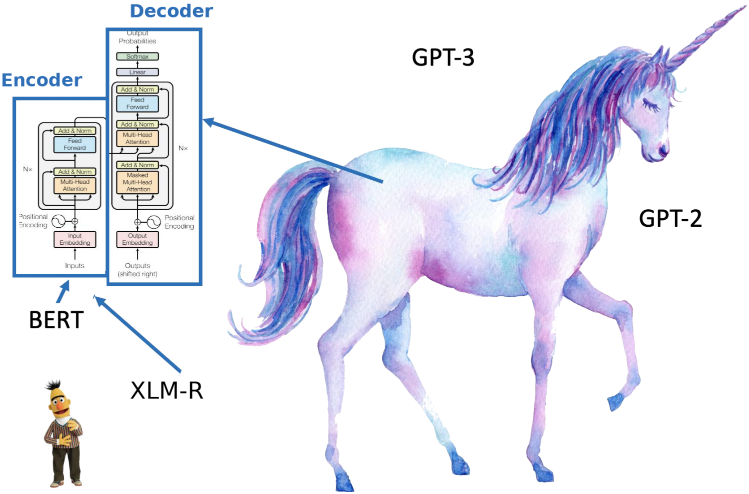
Embeddings, weights, parameters... Different words to say the same

Can they be initialised with pre-trained models?

(Vaswani et al., 2017)

Initialising (Multilingual) NMT

Copying the Weights: The Easy Way is not Easy



Decoder

Encoder

GPT-3

GPT-2

BERT

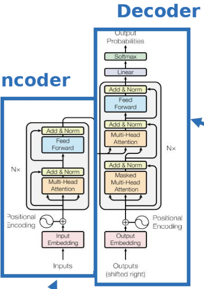
XLM-R

(Adapted from https://www.programmersought.com/article/24793362644/)

Initialising (Multilingual) NMT

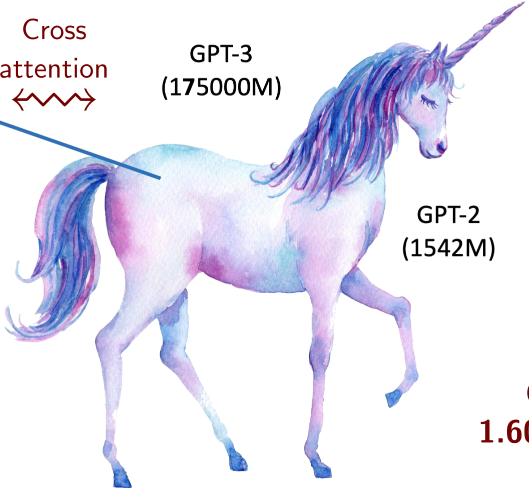
Copying the Weights: The Easy Way is not Easy

Number of layers



Cross attention

GPT-3 (175000M)



GPT-2 (1542M)

BERT (340M)

XLM-R (550M)

Google 1.600.000M!

(Adapted from https://www.programmersought.com/article/24793362644/)

Initialising (Multilingual) NMT

Copying the Weights: The Easy Way is not Easy

- It would be cool to be able to use embeddings from LMs trained with huge amount of data during weeks in powerful machines
- But pre-trained architectures are not supervised NMT friendly
- One can adapt NMT to match the LMs architectures
  (He et al.2018, Zhang et al.2020)

Initialising (Multilingual) NMT

Copying the Weights: The Easy Way is not Easy

- One can adapt NMT to match the LMs architectures (He et al. 2018, Zhang et al. 2020)
- One can train the LMs to mimic NMT blocks (Lample et al. 2019)
- One can do knowledge distillation to match the blocks (Chen et al. 2020)
- One can...

Initialising (Multilingual) NMT

Cross-lingual Language Model Pretraining (Lample & Conneau 2019)

- Train transformer with “NMT sizes” with monolingual corpora concatenated and CLM/MLM losses
- Initialise encoder and decoder, ignore cross-attention
- Ramachandran et al. 2016: for regularisation one should fine-tune with CLM/MLM + MT losses:
  - Some works cannot find improvements for other language pairs
  - catastrophic forgetting with different domain corpora

Initialising (Multilingual) NMT

Cross-lingual Language Model Pretraining (Lample & Conneau 2019)

	-		CLM		MLM	
	en-ro	ro-en	en-ro	ro-en	en-ro	ro-en
Sennrich 2016, BT	-	33.9	-	-	-	-
en → ro	28.6	-	31.0	-	36.3	-
ro → en	-	28.4	-	31.5	-	35.3
en ↔ ro	28.5	28.5	30.7	31.5	35.7	35.6
en ↔ ro + BT	35.9	34.4	37.8	37.0	39.1	38.5
Zhu 2020, Fusion	-	39.1	-	-	-	-

Results on supervised MT. BLEU scores on WMT'16 Romanian-English. The previous state-of-the-art of Sennrich 2016 uses both back-translation and an ensemble model. ro ↔ en corresponds to models trained on both directions.

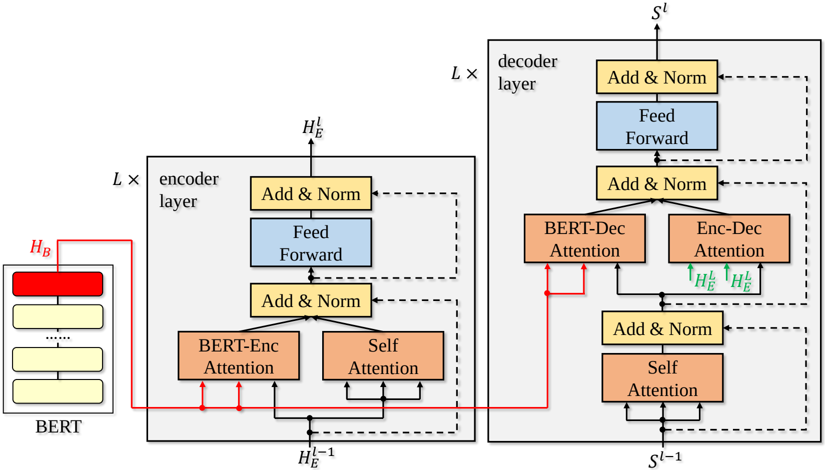
BERT in NMT, Fusion

Incorporating BERT into Neural Machine Translation (Zhu et al. 2020)

- Use BERT as it is; train an NMT
- Initialise BERT-fuse with the previous
- BERT is fused in each layer of the encoder and decoder of the NMT model using cross attention
- Drop-net probability decides how much BERT and how much NMT encoder and decoder to use

BERT in NMT, Fusion

Incorporating BERT into Neural Machine Translation (Zhu et al. 2020)



BERT in NMT, Fusion

Incorporating BERT into Neural Machine Translation (Zhu et al. 2020)

Algorithm	BLEU score
Standard Transformer	28.57
Use BERT to initialize the encoder of NMT	27.14
Use XLM to initialize the encoder of NMT	28.22
Use XLM to initialize the decoder of NMT	26.13
Use XLM to initialize both the encoder and decoder of NMT	28.99
Leveraging the output of BERT as embeddings	29.67

Preliminary explorations on IWSLT'14 English-to-German translation

BERT in NMT, Fusion

Incorporating BERT into Neural Machine Translation (Zhu et al. 2020)

	Transformer	BERT-fused
En2De	28.6	30.4
De2En	34.6	36.1
En2Es	39.0	41.4
En2Zh	26.3	28.2
En2Fr	35.9	38.7

BLEU of all IWSLT tasks

BERT in NMT, Fusion

Incorporating BERT into Neural Machine Translation (Zhu et al. 2020)

Standard Transformer	28.57
BERT-fused model	30.45
Randomly initialize encoder/decoder of BERT-fused model	27.03
Jointly tune BERT and encoder/decoder of BERT-fused model	28.87
Feed BERT feature into all layers without attention	29.61
Replace BERT output with random vectors	28.91
Replace BERT with the encoder of another Transformer model	28.99
Remove BERT-encoder attention	29.87
Remove BERT-decoder attention	29.90

Ablation study on IWSLT'14 English-to-German

Are we there?

Already at the End of the Way!

Thanks! And...

wait!

Thanks! And...

The List of Selected References

General: transformer, BERT, summary
[LLS20, VSP$^+$17, DCLT19]

Multilingual Embeddings: LASER
[AS19a, AS19b, LM20]

Multilingual Knowledge Distillation
[RG19, RG20]

Interlingual NMT Embeddings & SS-NMT
[EBBC17, EVBvG17, EvG18, REBvG19, EBR19, RvGE20]

Thanks! And...

References I

Mikel Artetxe and Holger Schwenk.
Margin-based parallel corpus mining with multilingual sentence embeddings.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy, July 2019. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk.
Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond.
Transactions of the Association for Computational Linguistics, 7(0):597–610, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
BERT: Pre-training of deep bidirectional transformers for language understanding.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Cristina España-Bonet and Alberto Barrón-Cedeño.
Lump at SemEval-2017 task 1: Towards an interlingua semantic similarity.
In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 144–149, Vancouver, Canada, August 2017. Association for Computational Linguistics.

Thanks! And...

References II

Cristina España-Bonet and Dana Ruiter.
UdS-DFKI participation at WMT 2019: Low-resource (en-gu) and coreference-aware (en-de) systems.
In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 183–190, Florence, Italy, August 2019. Association for Computational Linguistics.

Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith.
An empirical analysis of nmt-derived interlingual embeddings and their use in parallel sentence identification.
IEEE Journal of Selected Topics in Signal Processing, 11(8):1340–1350, December 2017.

Cristina España-Bonet and Josef van Genabith.
Multilingual Semantic Networks for Data-driven Interlingua Seq2Seq Systems.
In *Proceedings of the LREC 2018 MLP-MomenT Workshop*, pages 8–13, Miyazaki, Japan, May 2018.

Zhiyuan Liu, Yankai Lin, and Maosong Sun.
Sentence representation.
10.1007/978-981-15-5573-2_4., 2020.

Wei Li and Brian Mak.
Transformer based Multilingual document Embedding model.
arXiv e-prints, page arXiv:2008.08567, August 2020.

Thanks! And...

References III

Dana Ruiter, Cristina España-Bonet, and Josef van Genabith.
Self-Supervised Neural Machine Translation.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers.*, pages 1828–1834, Florence, Italy, August 2019. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych.
Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.
In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics, 2019.

Nils Reimers and Iryna Gurevych.
Making monolingual sentence embeddings multilingual using knowledge distillation.
In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4512–4525. Association for Computational Linguistics, 2020.

Thanks! And...

References IV

Dana Ruiter, Josef van Genabith, and Cristina España-Bonet.
Self-Induced Curriculum Learning in Self-Supervised Neural Machine Translation.
In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2560–2571, Online, November 2020. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.

Multilingual Sentence Embeddings in/and/for Neural Machine Translation

Cristina España-Bonet
DFKI GmbH

Recent Advances in Machine Translation (RAMT 2021)

Webex, everywhere on the Earth
(with internet)
18th March 2021