An Evaluation of Language- Agnostic Inner-Attention-Based Representations in Machine Translation

Alessandro Raganato **Raúl Vázquez** **Mathias Creutz** **Jörg Tiedemann**
name.lastname@helsinki.fi

FOTRAN Found in Translation

UNIVERSITY OF HELSINKI

ACADEMY OF FINLAND

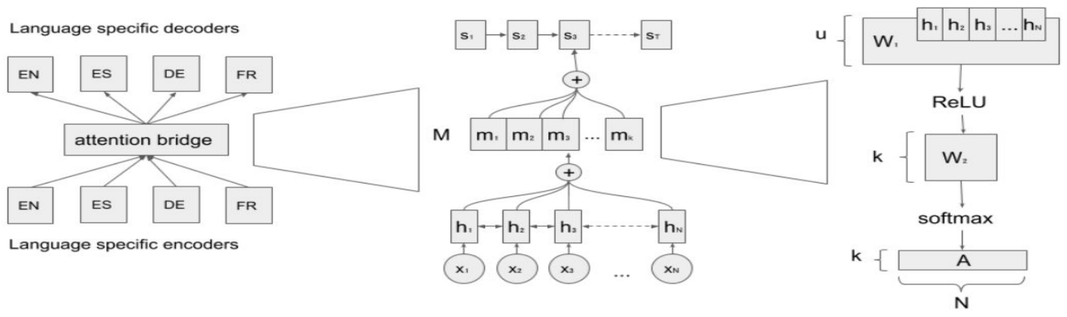
European Research Council
Established by the European Commission

Study

exploring a multilingual translation model with a fixed size shared layer that can be used as sentence representation in different downstream tasks.

systematically study the impact of the size of the shared layer and the effect of including additional languages in the model, both in translation quality and in classification tasks.

Model Architecture



- Standard encoder-decoder model with traditional attention mechanism.
- Language-specific encoders/decoders enable multilingual training
- Shared inner-attention layer embeds sentences into fixed-size vector space
- Each decoder receives information only through the shared attention bridge:
  - context-vector are computed by attending to the states of the attention bridge;
  - decoders are initialized by mean pooling over the shared layer.

Experimental Setup

Models specifications:

Embedding layers: 512 dimensions,

Encoders: two biLSTMs with 512 hidden units,

Decoders: two LSTMs with 512 units & traditional attention

Language scheduler: uniformly distributed

Attention Bridge: 1, 10, 25 and 50 attention heads (*k*) with 1024 hidden units each

Languages:
English (EN)
French (FR)
Spanish (ES)
German (DE)

Datasets:

Train: ~2M sentences per language pair (Europarl v7)

Dev: 2K sentences from dev2006 (ACL-WMT07)

Test: 4K sentences from devtest+test2006 (ACL-WMT07)

We trained

- bilingual models for EN→DE;
- multilingual models {DE,ES,FR }↔ EN
- Many-to-Many model (only 50 *k*)

- Best model optimizes BLEU score on the validation set
- Evaluate sentence representations using the SentEval toolkit
- Evaluate models translation quality

SentEval

Classification Tasks

		SNLI	SICK-E	AVG
en→de	k=1	63.86	77.09	71.46
en→de	k=10	**65.30**	78.77	72.02
en→de	k=25	65.13	79.34	**72.68**
en→de	k=50	**65.30**	**79.36**	72.60
Multilingual	k=1	65.56	77.96	72.67
Multilingual	k=10	67.01	79.48	72.89
Multilingual	k=25	66.94	79.85	**73.67**
Multilingual	k=50	**67.38**	**80.54**	73.39
Many-to-Many	k=50	**67.73**	**81.12**	**74.33**
Most frequent baseline†		34.30	56.70	48.19
GloVe-BOW†		66.00	78.20	75.81
Cífka and Bojar (2018) en→cs†		69.30	80.80	73.40

Table 1: Accuracy of different models on two SentEval tasks as well as the overall average accuracy on all of them. The general trend is that a higher number of attention heads and multilingual models are beneficial. Results with † taken from Cífka and Bojar (2018).

- general trend:
  - ➢ accuracies improve with larger representations.
- Positive effect of multilingual training:
  - ➢ Manyto-Many model performs best on average even though it does not add any further training examples for English (compared to the other multilingual models), which is the target language of the downstream tasks.

Similarity Tasks

		SICK-R	STSB	AVG
en→de	k=1	0.74 / 0.67	**0.69 / 0.69**	**0.57**
en→de	k=10	0.76 / 0.71	**0.69 / 0.69**	0.52
en→de	k=25	**0.78 / 0.73**	0.67 / 0.66	0.49
en→de	k=50	0.78 / 0.72	0.65 / 0.64	0.46
Multilingual	k=1	0.76 / 0.71	0.69 / 0.68	**0.50**
Multilingual	k=10	0.78 / 0.74	**0.69 / 0.69**	0.48
Multilingual	k=25	0.78 / 0.74	0.68 / 0.67	0.43
Multilingual	k=50	**0.79 / 0.74**	0.66 / 0.64	0.40
Many-to-Many	k=50	**0.79 / 0.74**	**0.69 / 0.68**	0.40
InferSent†		0.88 / 0.83	0.76 / 0.75	0.66
GloVe-BOW†		0.80 / 0.72	0.64 / 0.62	0.53
Cífka and Bojar (2018) en→cs†		0.81 / 0.76	0.73 / 0.73	0.45

Table 2: Results from supervised similarity tasks (SICK-R and STSB), measured using Pearson's (r) and Spearman's (ρ) correlation coefficients (r/ρ). The average across unsupervised similarity tasks on Pearson's measures are displayed in the right-most column. Results with † taken from Cífka and Bojar (2018).

- on the unsupervised textual similarity tasks, having fewer attention heads is beneficial while negative effect of multilingual models
- On the supervised textual similarity tasks, higher number of attention heads and multilinguality contribute to better scores

Translation Quality

		k=1	k=10	k=25	k=50	M-to-M	att.
en	de	14.66	19.87	20.61	20.83	20.47	22.72
	es	21.82	27.55	28.41	28.13	27.6	30.28
	fr	17.8	23.35	24.36	23.79	24.15	25.88
de	en	16.97	21.39	23.42	24	24.4	24.28
es		18.38	25.39	27.01	27.12	26.98	28.16
fr		17.52	21.93	24.4	23.9	24.47	25.39

Table 3: BLEU scores for multilingual models. Baseline system in the right-most column.

- more attention heads lead to a higher BLEU score
- the model with 50 heads achieves the best results
- same ballpark in translation quality from the {DE,ES,FR}↔EN model and the Many-to-Many model

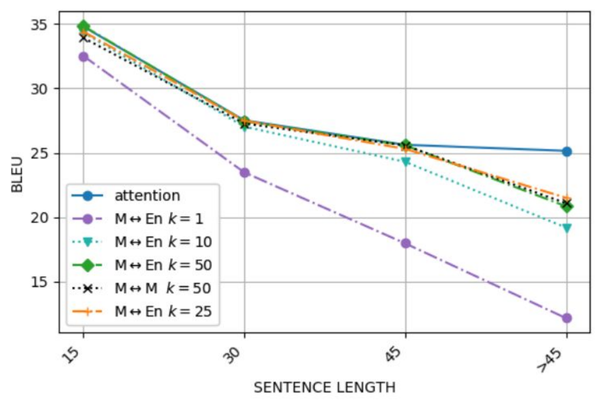


Figure 2: The BLEU scores obtained by the multilingual models and baseline system with respect to different sentence length.

- a larger number of attention heads has a positive impact when translating longer sentences.
- the performance drop of the attention bridge models is entirely due to sentences longer than 45 words.