Multilingual neural machine translation (NMT) with a language-independent attention bridge

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

FOTRAN Found in Translation

UNIVERSITY OF HELSINKI

ACADEMY OF FINLAND

European Research Council
Established by the European Commission

Study

We propose an architecture for multilingual machine translation (MT) capable of obtaining multilingual sentence representations by means of incorporating an intermediate cross-lingual shared layer (*attention bridge*) in multilingual training.

- Exploits the semantics from each language and develops into a language-agnostic meaning representation
- Can efficiently be used for transfer learning.
- Encoder-decoder model with three important additions;
  1. Attention bridge

$$A = softmax\left(W_2 \mathrm{ReLU}(W_1 H^T)\right)$$
$$M = AH$$

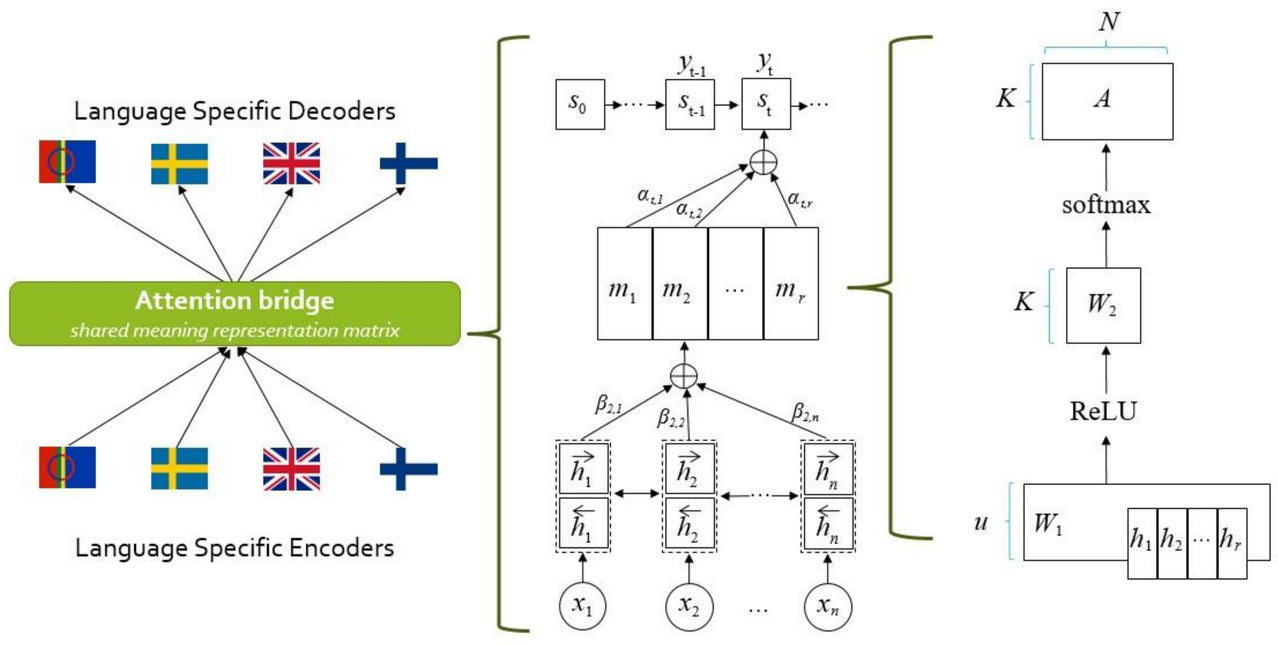
  where W_1 and W_2 are weight matrices and H are the hidden states of the encoder

  2. Penalty term in the loss function

$$\mathcal{L} = -\log\left(p\left(Y|X\right)\right) + \left\|AA^T - I\right\|_F^2$$

  where I is the identity matrix. Helps avoid repetitive information

  3. Language-specific encoders and decoders



Proposed multilingual NMT system: (left) the attention bridge connects the language-specific encoders and decoders; (center) an overview of the model for one language pair; (right) computation of the fixed-size attentive matrix A, used to obtain the attention bridge.

Experimental Setup

Trained on the multi30K dataset:

29,000 sentences for train
1,000 sentences for dev
1,000 sentences for test
from flickr2016

Languages:
English (En)
French (Fr)
Czech (Cs)
German (De)

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: 10 attention heads & 1024 hidden units

Focus on multilingual transfer learning in low-resource
1. Evaluation of translation quality and zero-shot
2. Test of the produced sentence encoding by using them in the Senteval tasks
3. Study the effect of the penalty term

Multilingual Translation of Image Captions

The attention bridge effectively encodes and shares multilingual information across various language pairs.

- Baseline: verify the correct functionality of the architecture in a bilingual setting
- {De,Fr,Cs}↔En: Multilingual setting outperforms bilingual baselines for language-pairs seen during training
  Including monolingual data leads to increasing the BLEU scores & enables zero-shot translation
- Many2Many: training with monolingual data leads to the overall best model.
  Improvements in BLEU range from 1.40 to 4.43 when compared to the standard bilingual model.

	BILINGUAL				{DE,FR,CS} ↔ EN				M-2-M			
src/tgt	EN	DE	CS	FR	EN	DE	CS	FR	EN	DE	CS	FR
EN	-	36.78	28.00	55.96	-	37.85	29.51	57.87	-	37.70	29.67	55.78
DE	39.00	-	23.44	38.22	39.39	-	*0.35*	*0.83*	40.68	-	26.78	41.07
CS	35.89	28.98	-	36.44	37.20	*0.65*	-	*1.02*	38.42	31.07	-	40.27
FR	49.54	32.92	25.98	-	48.49	*0.60*	*0.30*	-	49.92	34.63	26.92	-
	BILINGUAL + ATT BRIDGE				{DE,FR,CS} ↔ EN + MONOLING				M-2-M + MONOLINGUAL			
	EN	DE	CS	FR	EN	DE	CS	FR	EN	DE	CS	FR
EN	-	35.85	27.10	53.03	-	38.92	30.27	57.87	-	38.48	30.47	57.35
DE	38.19	-	23.97	37.40	40.17	-	*19.50*	*26.46*	41.82	-	26.90	41.49
CS	36.41	27.28	-	36.41	37.30	*22.13*	-	*22.80*	39.58	31.51	-	40.87
FR	48.93	31.70	25.96	-	50.41	*25.96*	*20.09*	-	50.94	35.25	28.80	-

Table 1: BLEU scores obtained in the experiments. *Left:* Bilingual models, our baselines. *Center:* Models trained on {De,Fr,Cs}↔En, with zero-shot translations in italics. *Right:* Many-to-many model. Both zero-shot and M-2-M translations improve significantly when including monolingual data. (Best results in green cells.)

SentEval

The multilingual models sentence embeddings get better results than their bilingual counterparts

Classification Tasks: Our Many2Many model obtains better results in the trainable semantic similarity tasks. (SICK Relatedness and STS-Benchmark).

Probing Tasks: Significant increment on the the specific tasks of Length *(superficial property)*, Top Constituents *(syntactic property)* and Object Number *(semantic information)*

Multilingual models outperform the bilingual models in all but one test.

	DOWNSTREAM TASKS			
TASK	BASELINE	M ↔ EN	M-2-M	GloVe-BoW
CR	68.52	68.32	69.01	63.97
MR	60.08	60.40	61.80	52.32
MPQA	73.51	72.98	73.28	68.76
SUBJ	77.25	78.64	80.88	58.75
SST2	61.92	62.02	62.24	54.68
SST5	31.15	32.10	31.83	28.20
TREC	67.75	69.84	66.40	21.16
MRPC	70.96	68.83	70.43	64.87
SNLI	61.75	64.52	65.12	35.05
SICKE	74.85	75.46	76.92	56.62
SICKR	*0.652*	*0.659*	*0.677*	*0.174*
STS-B	*0.616*	*0.618*	*0.630*	*0.163*
PROBING TASKS				
Length	80.76	84.76	85.41	30.90
WC	10.02	9.56	9.13	0.22
Depth	32.14	33.05	31.60	20.66
TopConst	40.12	44.04	39.76	11.48
BShift	57.41	58.35	59.76	50.08
Tense	67.61	69.36	68.27	54.72
SubjNum	68.55	69.67	69.89	54.32
ObjNum	70.01	72.19	73.29	60.58
SOMO	49.90	49.46	50.12	50.03
CoordInv	61.38	60.57	62.21	49.88

Table 2: Scores obtained in the SentEval tasks. The BASELINE column reports the best score among the bilingual models + att bridge. Green cells indicate the highest score. All tasks show the accuracy of the model except for SICKR and STS-B tasks, which include Pearson mean values.

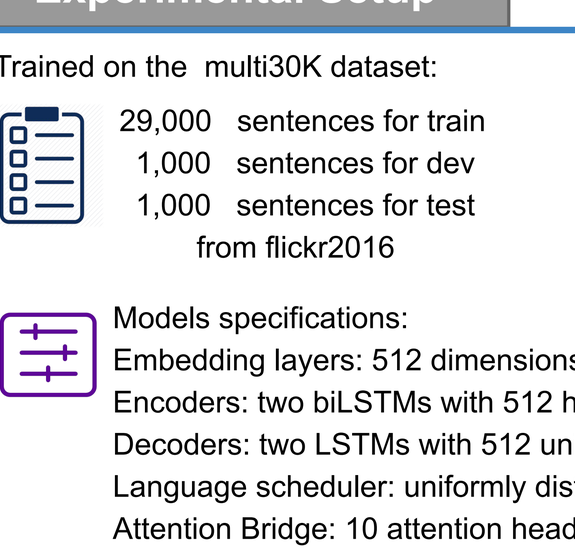


Figure 1: For every language pair, we compare the BLEU scores between our best model (M-2-M with monolingual data), the zero-shot of the model trained without that specific language pair and the bilingual model of that language pair.

We analyze the zero-shot translation capabilities in more detail.

Train six different models where we include all but one of the available language pairs.

	WITH PENALTY TERM			
	EN	DE	CS	FR
EN	-	35.85	27.10	53.03
DE	38.19	-	23.97	37.40
CS	36.41	27.28	-	36.41
FR	48.93	31.70	25.96	-
	WITHOUT PENALTY TERM			
	EN	DE	CS	FR
EN	-	34.67	27.22	54.39
DE	38.70	-	23.44	38.2
CS	35.76	28.50	-	36.4
FR	48.76	31.60	25.55	-

Table 3: BLEU scores obtained with the BILINGUAL + ATT BRIDGE models in the experiments with and without penalty term.

The penalty term encourages the attentive matrix to focus on different aspects of the sentence