HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Multilingual NMT with a language-independent attention bridge

Raúl Vázquez

November 2018

Department of Digital Humanities

FOTRAN
Found in Translation

Multilingual Neural Machine Translation

What?

- MT that translates between multiple languages
- 3 strategies:
  - one-to-many
  - many-to-one
  - many-to-many

Why?

- Better translations for low-resourced languages
- Enables *zero-shot translation*

Sentence Representations

- Fixed-size sentence representations embedded in continuous vector spaces.
- Useful:
  - testing downstream tasks
  - enable a deeper linguistic analysis
  - better understanding what the neural models are learning
- Seq2Seq NMT models (Sutskever et al., 2014) have a natural way of generating sentence representations
- Replaced by the use of attention mechanisms (Bahdanau et al., 2014)

$\therefore$ we want a model s.t.

1. produces good quality translations
2. efficiently uses transfer learning
3. produces a fixed size sentence embedding

$\therefore$ we want a model s.t.

1. produces good quality translations $\longrightarrow$ obvious benefits ;)
2. efficiently uses transfer learning $\longrightarrow$ especially useful for low-resource scenarios
3. produces a fixed size sentence embedding $\longrightarrow$ would allow for probing and downstream testing tasks

Proposed Model

Hence, we propose the following multilingual MT model:

- An attention based encoder-decoder architecture with 3 modifications:

  (i) a shared self-attention layer (*the attention bridge*)

  (ii) language-specific encoders and decoders

  (iii) a penalty term in the loss function

FOTRAN
Found in Translation

Background

Attention Mechanism

Given an input $X = (x_1, \ldots, x_n) \in \mathbb{R}^{d_x}$ generate a translation $Y = (y_1, \ldots, y_m)$.
Encoder: an RNN that genertes a contex vector c from X. Generally:

$$h_t = f(x_t, h_{t-1}) \ ; \qquad c = h_n \tag{1}$$

with $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_h} \longrightarrow \mathbb{R}^{d_h}$ a non-linear activation function. We use bidirectional LSTM units.
Decoder: sequentially computes $(y_1, \ldots, y_m)$ by optimizing

$$p(Y|X) = \prod_{t=1}^{m} p(y_t|c, Y_{t-1}) \ ; \qquad Y_{t-1} = (y_1, \ldots, y_{t-1}) \tag{2}$$

Each distribution $p_t = p(y_t|c,) \in \mathbb{R}^{d_v}$ is usually computed with a softmax function over the vocabulary:

$$p_t = softmax(y_{t-1}, s_t) \ ; \qquad s_t = \varphi(c, y_{t-1}, h_{t-1}) \tag{3}$$

where φ is another non-linear activation function and d_v is the size of the vocabulary.
Attention mechanism $\Longrightarrow$ a different context vector c_t will be computed at each step t. By defining $c_t = \sum_{i=1}^{n} \alpha_{t,i} h_t$, where $\alpha_{t,i}$ indicates how much the i-th input word contributes to generating the t-th output word,

$$\alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{k=1}^{n} exp(e_{t,k})} \ ; \qquad e_{t,i} = g(s_t, h_i) \tag{4}$$

and g is a feedforward neural network.

FoTran Found in Translation

Model Architecture

Background: Attention Mechanisms

For the purpose of this presentation:

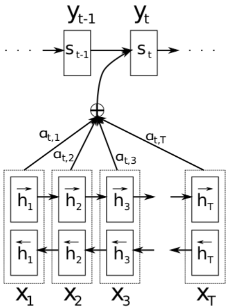


Figure 1: Alignment model proposed by Bahdanau et al. (2014)

An extension of the attention-based model with 3 modifications:

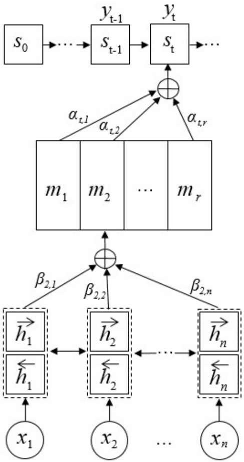
(i) the attention bridge

(ii) language-specific encoders and decoders

(iii) a penalty term in the loss function

NOTE: *the architecture is not restricted to RNN-based encoders/decoders*

(i) the attention bridge:



- Encodes fixed-size (language-independent) sentence representations.
- Can attent r different components of the sentence.
- Embeds the hidden states $H = (h_t) \in \mathbb{R}^{d_h \times X_T}$ into a fixed size matrix $M \in \mathbb{R}^{d_h \times r}$

$$B = softmax\,(W_2\mathrm{ReLU}(W_1 H))$$

$$M = BH^T$$

- Compound attention model (Cífka and Bojar, 2018)

(ii) language-specific encoders and decoders

- one NN encoder for each input language.
- one attentive decoder for each output language.
- trainable with a language scheduler.
- neural-interlingua (Lu et al., 2018)

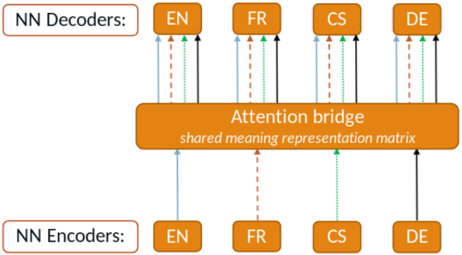


Figure 2: Multiple encoders/decoders with an additional self-attention layer

(ii) language-specific encoders and decoders

- one NN encoder for each input language.
- one attentive decoder for each output language.
- trainable with a language scheduler.
- neural-interlingua (Lu et al., 2018)

NOTE: *the architecture is not restricted to RNN-based encoders/decoders*

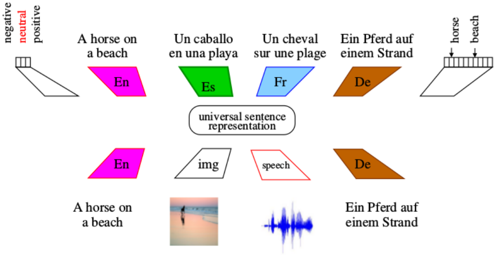


Figure 3: Generic multilinfual and -modal encoder--decoder architecture (Schwenk and Douze, 2017)

(iii) penalty term

We want the attention bridge layer to illustrate various components of a sentence

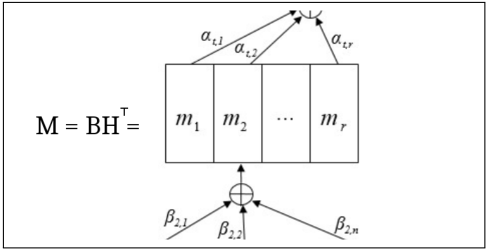


Figure 4: Zoom of the attention bridge in the compound architecture

- Matrix M could learn repetitive information
- We use the loss function:

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

- The penalty forces matrix $BB^T \sim I$

Looks like a nice idea! ...

So, how well does it perform?

The multi30k models

Dataset

- Multi-parallel dataset of image captions
- Languages: En, De, Cs, Fr
- 29k captions for training
- Tested on 1k captions from flickr 2016 testset

The multi30k models

Dataset

- Multi-parallel dataset of image captions
- Languages: En, De, Cs, Fr
- 29k captions for training
- Tested on 1k captions from flickr 2016 testset

Hyperparameters:

- 10k BPE $\times$ language
- 1 encoder per language:
  2 stacked BiLSTMs of size $d_h = 512$
- 1 decoder per language:
  2 attentive LSTMs of size $d_h = 512$
- The attention bridge:
  10 attention heads
  each of dimension 512

FOTRAN Found in Translation

The multi30k models

Dataset

- Multi-parallel dataset of image captions
- Languages: En, De, Cs, Fr
- 29k captions for training
- Tested on 1k captions from flickr 2016 testset

* *We implemented our model on our OpenNMT-py fork*

https://github.com/Helsinki-NLP/OpenNMT-py/tree/neural-interlingua

Hyperparameters:

- 10k BPE $\times$ language
- 1 encoder per language:
  2 stacked BiLSTMs of size $d_h = 512$
- 1 decoder per language:
  2 attentive LSTMs of size $d_h = 512$
- The attention bridge:
  10 attention heads
  each of dimension 512

Baselines

	EN	DE	CS	FR
BILINGUAL				
EN	-	36.78	28.00	55.96
DE	39.00	-	23.44	38.22
CS	35.89	28.98	-	36.44
FR	49.54	32.92	25.98	-
BILINGUAL + ATTENTION BRIDGE				
	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	-

Table 1: 24 bilingual baseline models BLEU. All models share specifications, apart from the proposed changes to include the attention bridge layer for the second part of the table.

- Examine performance in a bilingual setting
- Slight drop in performance due to the fixed-size attention bridge
- Architecture robust enough for translation

Many-To-One and One-To-Many Models

{DE,FR,CS} ↔ EN				
	EN	DE	CS	FR
EN	-	37.85	29.51	57.87
DE	39.39	-	0.35	0.83
CS	37.20	0.65	-	1.02
FR	48.49	0.60	0.30	-
{DE,FR,CS} ↔ EN + MONOLINGUAL				
	EN	DE	CS	FR
EN	-	38.92	30.27	57.87
DE	40.17	-	19.50	26.46
CS	37.30	22.13	-	22.80
FR	50.41	25.96	20.09	-

Table 2: BLEU scores obtained for models trained on {De,Fr,Cs}↔En. Zero-shot translation marked by the shaded cells.

- The power of the attention bridge:
  share information across various languages
- Seen language pairs are boosted
- Zero-shot translation only when including monolingual data during training.
- This boosts the seen language pairs scores.

Many-to-Many Models

M-2-M	EN	DE	CS	FR
EN	-	37.70	29.67	55.78
DE	40.68	-	26.78	41.07
CS	38.42	31.07	-	40.27
FR	49.92	34.63	26.92	-

M-2-M + MONOLINGUAL	EN	DE	CS	FR
EN	-	38.48	30.47	57.35
DE	41.82	-	26.90	41.49
CS	39.58	31.51	-	40.87
FR	50.94	35.25	28.80	-

Table 3: The multilingual model also gets a boost when incorporating monolingual data during training.

- More language pairs ⇒ better performance.
- Seen language pairs are boosted
- Including monolingual data during training boosts the seen language pairs scores.
- This produces the overall best model trained on multi30k

FOTRAN Found in Translation

SentEval-multi30k

SentEval					
TASK	en-de	en-cs	en-fr	m<->en	m2m
MR	59.52	58.75	59.34	60.13	**61.65**
SUBJ	74.97	75.82	76.18	78.73	**80.39**
SST2	62.16	62.55	62.88	**64.03**	62.22
SST5	30.41	31.09	31.81	**32.49**	30.14
TREC	70.8	65	63	**71.2**	62.4
MRPC	68.52	67.88	69.04	69.04	**70.72**
SICKEntailment	73.17	**77.2**	74.69	74.73	76.86
Length	64.28	67.8	66.78	74.01	**75.55**
WordContent	28.17	**28.89**	25.63	24.85	21.51
Depth	29.75	29.06	30.05	31.47	**31.8**
TopConstituents	52.38	52.88	50.7	**58.73**	51.97
BigramShift	55.44	54.81	55.05	56.93	**57.25**
Tense	66.62	65.3	68.81	74.28	**75.57**
SubjNumber	65.89	63.74	69.07	**71.87**	71.02
ObjNumber	65.55	65.96	70.34	73.86	**76.01**
OddManOut	49.58	49.33	**50.55**	49.68	49.92
CoordinationInversion	56.69	56.59	56.87	**58.4**	57.65

Figure 5: muli30k models SentEval evaluation. Tasks that report accuracy.

Looks like it is doing the trick!

How about a bigger dataset?

The europarl models

Dataset

- Non multi-parallel dataset
- from the Proceedings of the European Parliament
- Languages: En, De, Es, Fr
- Training:

En-De $\sim$ 1M parallel sentences
En-Es " " " "
En-Fr " " " "

- not tested yet

The europarl models

Dataset

- Non multi-parallel dataset
- from the Proceedings of the European Parliament
- Languages: En, De, Es, Fr
- Training:

En-De $\sim$ 1M parallel sentences
En-Es " " " "
En-Fr " " " "

- not tested yet

Hyperparameters:

- 32k BPE $\times$ language
- 1 encoder per language:
  2 stacked BiLSTMs of size $d_h = 512$
- 1 decoder per language:
  2 attentive LSTMs of size $d_h = 512$
- The attention bridge:
  10 attention heads
  each of dimension 512

FOTRAN Found in Translation

Europarl

BILINGUAL		
EN	DE	23.85
	ES	33.71
	FR	28.21
DE	EN	29.97
ES		34.74
FR		30.42

BILINGUAL + ATT.BRIDGE		
EN	DE	18.49
	ES	28.39
	FR	23.02
DE	EN	24.68
ES		28.91
FR		24.71

{DE,ES;FR} <-> EN + Monolingual		
EN	DE	19.08
	ES	28.71
	FR	23.08
DE	EN	24.64
ES		29.19
FR		27.67

Figure 6: europarl models BLEU score reported during validation.

SentEval-europarl

SentEval				
TASK	en-de	en-es	en-fr	m<->en
MR	66.93	67.6	67.13	**68.47**
SUBJ	85.43	85.48	85.79	**86.79**
SST2	71.94	69.58	71.99	**73.2**
SST5	36.47	37.24	37.47	**39.77**
TREC	75.8	**80.2**	76.2	76
MRPC	**74.09**	67.83	73.51	72.41
SICKEntailment	74.81	76.05	73.9	**76.27**
Length	82.39	82.57	81.39	**85.78**
WordContent	35.42	31.72	30.54	**50.96**
Depth	36.25	35.71	35.37	**36.62**
TopConstituents	66.33	67.88	67.13	**72.71**
BigramShift	59.7	60.61	59.9	**64.25**
Tense	82.37	82.35	82.58	**83.33**
SubjNumber	79.65	80.76	80.11	**81.8**
ObjNumber	79.26	82.48	80.69	**83.4**
OddManOut	**50.9**	49.79	49.78	50.36
CoordinationInversion	**60.8**	57.95	59.44	59.55

Figure 7: europarl models SentEval evaluation. Tasks that report accuracy.

FOTRAN Found in Translation

Conclusions

- We propose a multilingual NMT architecture - openly available to the public
- We develop a multilingual MT system that
  - efficiently incorporates transfer learning
  - can learn learning multilingual sentence representations.
- The inclusion of monolingual data during training resulted in boosted scores for all cases.

multi30k: multilingual models outperform their bilingual counterparts $\Rightarrow$ efficiently shares parameters

europarl: not really $\Rightarrow$ If one has enough data to train strong bilingual models, why bother to use multilingual?
BUT this can def. serve for domain adaptation towards other low-resourced languages.

Thank You!