Are BLEU and Meaning Representation in Opposition?

Ondřej Cífka
Ondřej Bojar

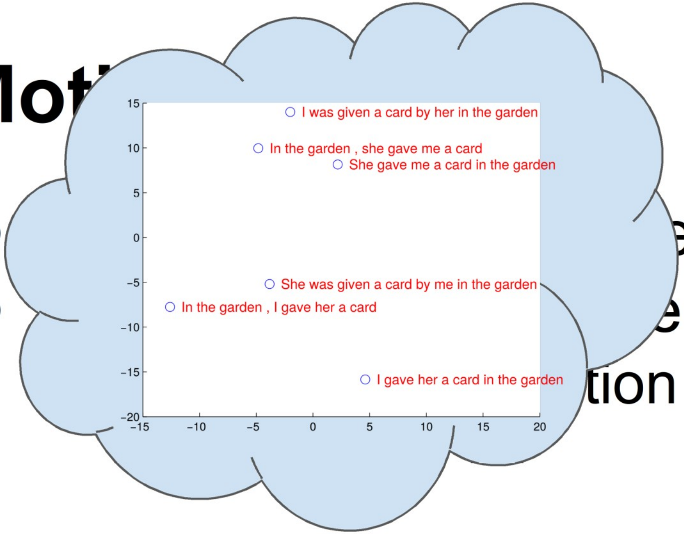
FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

ÚFAL

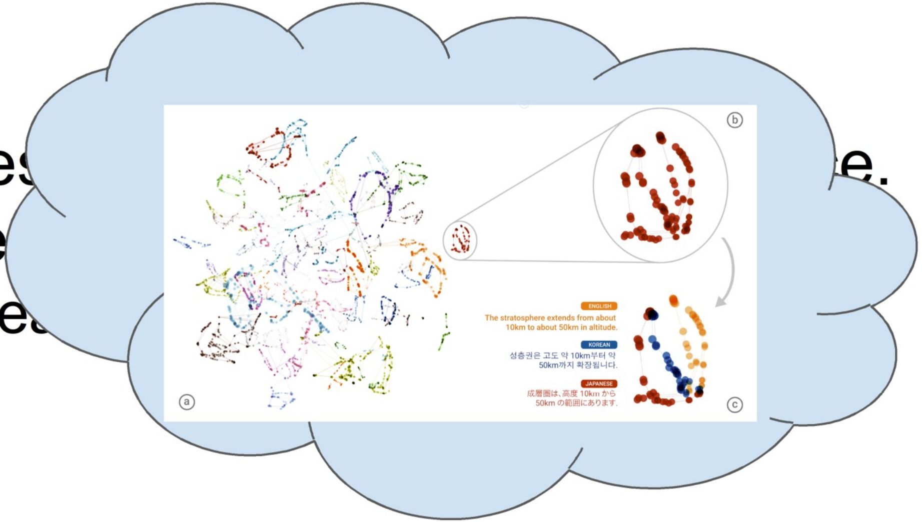
Motivation

- Good translation preserves the meaning of the sentence.
- Neural MT learns to represent the sentence.
    - Is the representation “meaningful” in some sense?

Mot

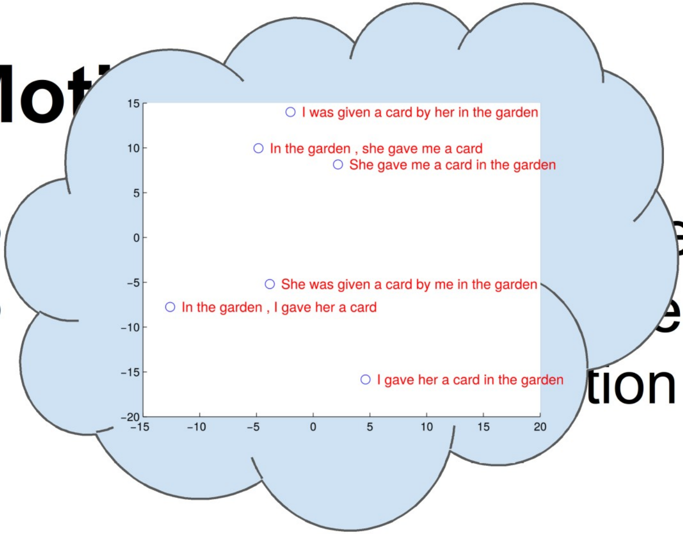


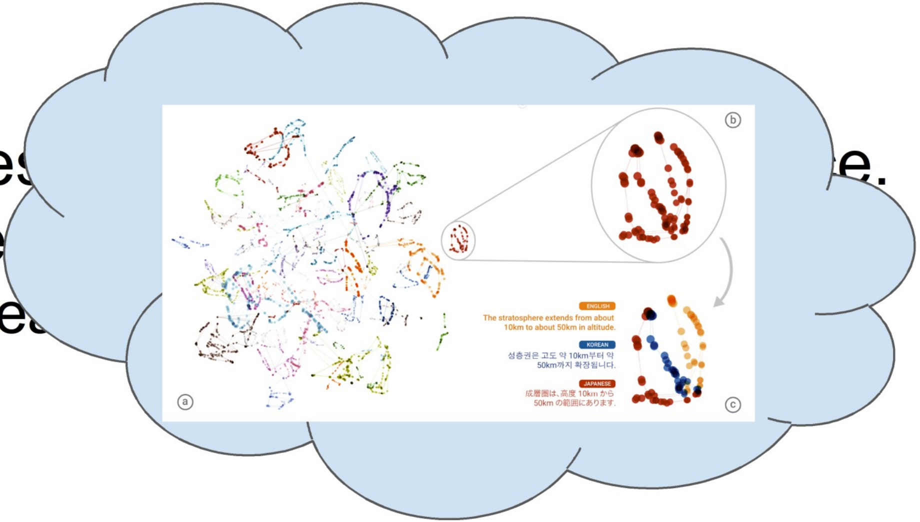
- erves … e.
- epre
  tion “mea



Mot

- erves e.
- epre
  tion “me





Gist of our idea:

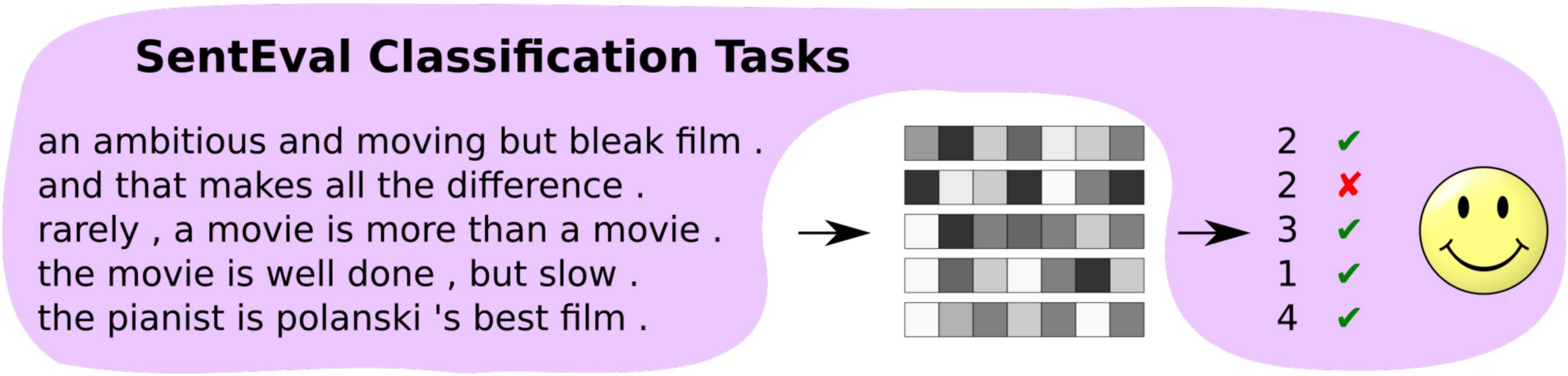
1. Train variants of NMT to obtain sentence representations.
2. Evaluate all such representations “semantically”.
3. Relate performance in MT and in “semantics”.

Evaluating sentence representations

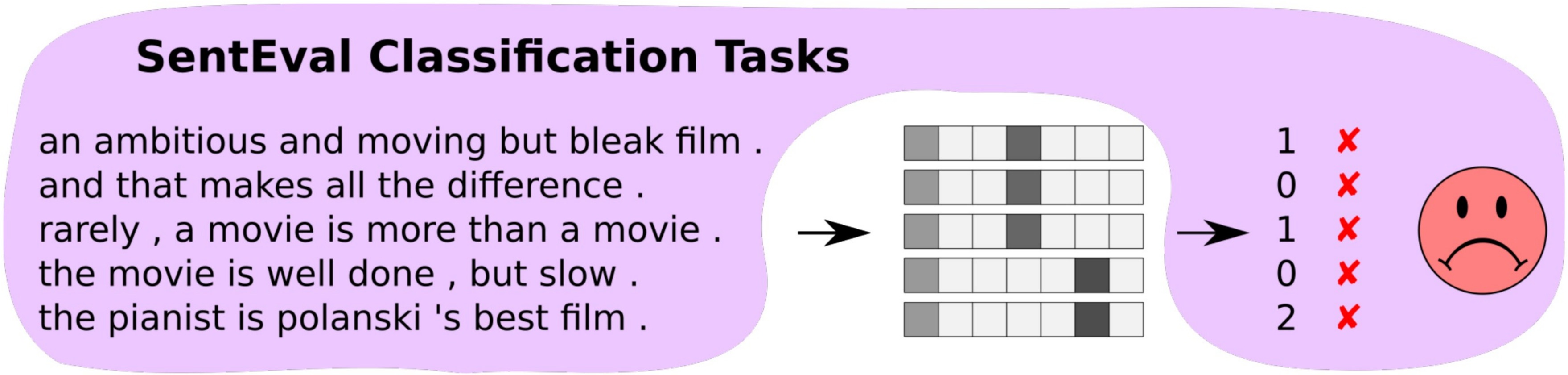
- Evaluation through classification.
- Evaluation through similarity.
- Evaluation using paraphrases.

- SentEval (Conneau et al., 2017)
  - prediction tasks for evaluating sentence embeddings
  - focus on semantics (recently, “linguistics” task added, too).
- HyTER paraphrases (Dreyer and Marcu, 2014)

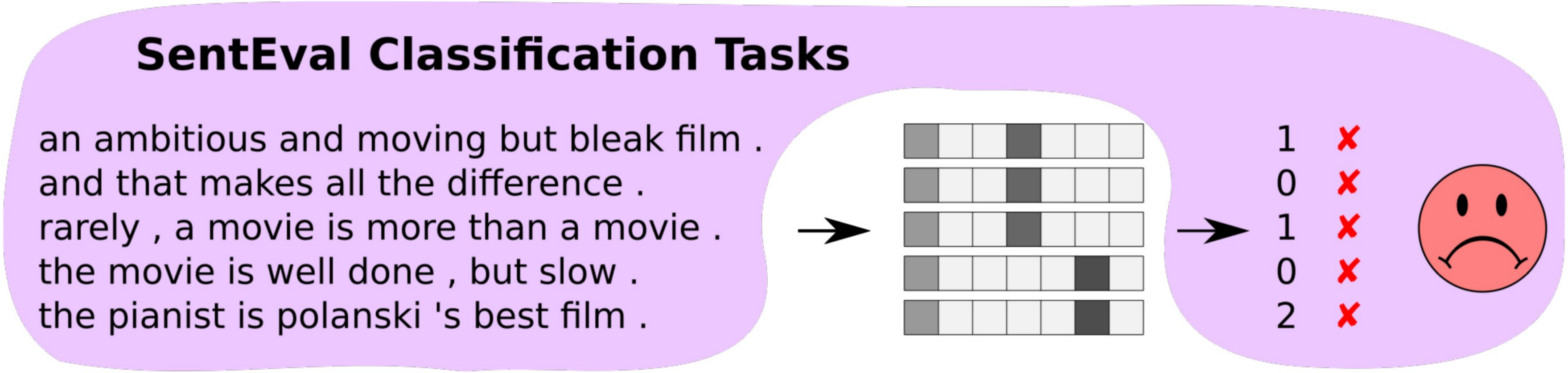
Evaluation through classification



Evaluation through classification

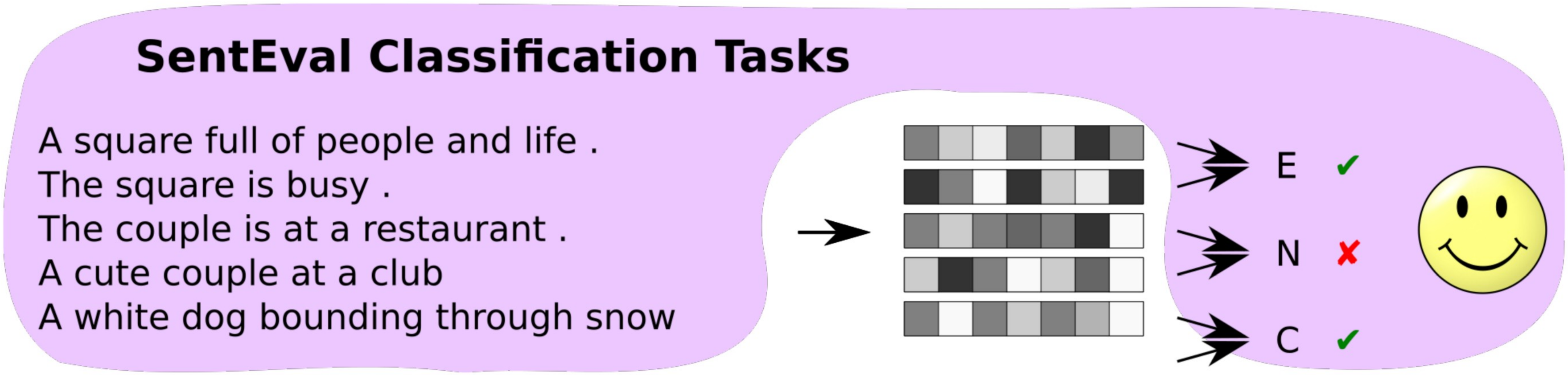


Evaluation through classification



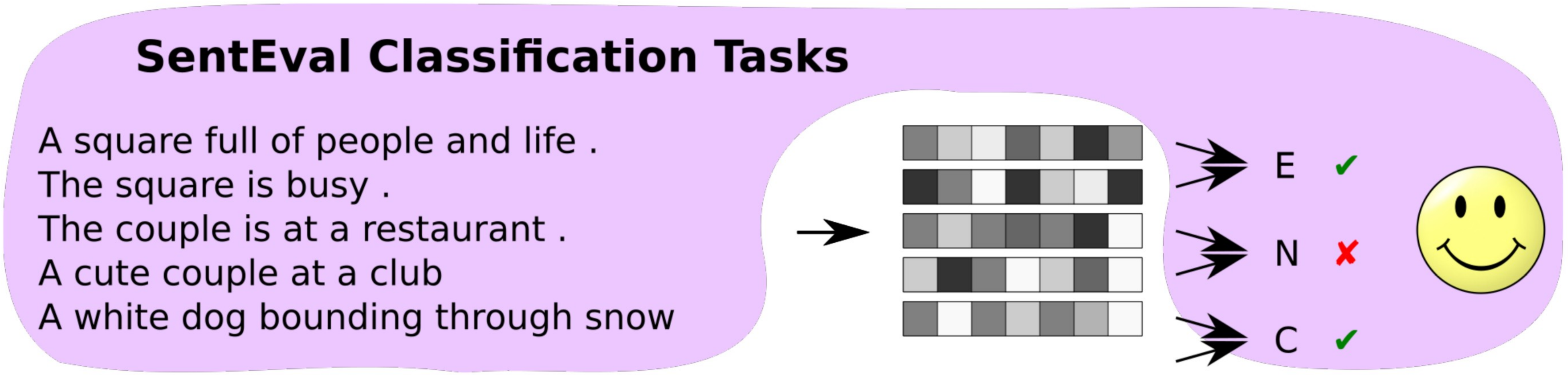
- Solo: movies sentiment, product review polarity, question type...

Evaluation through classification



- Solo: movies sentiment, product review polarity, question type...
- Paired: natural language inference, semantic equivalence

Evaluation through classification



- Solo: movies sentiment, product review polarity, question type...
- Paired: natural language inference, semantic equivalence
- 10 classification tasks in total, we report them as “AvgAcc”
  - 4k-55k training examples, with testset or 10-fold crosseval.

Evaluation through similarity

- 7 similarity tasks: pairs of sentences + human judgement

I think it probably depends on your money.	It depends on your country.	0
Yes, you should mention your experience.	Yes, you should make a resume	2
Hope this is what you are looking for.	Is this the kind of thing you're looking for?	4

  - with training set, sent. similarity predicted by regression,
  - without training set, cosine similarity used as sent. sim.,
  - ultimately, the predicted sent. similarity is correlated with the golden truth.
- In sum, we report them as “AvgSim”.

Evaluation using paraphrases: the data

- HyTER: ~200 sentences, 500 translations each
- COCO: 5k images, 5 captions each

低胸露背的黄金泳衣重五百公克，售价一千万日币。

the deep cut and halter golden swimwear weighs half kilogram selling at ten million JPY.

¥10,000,000 is the retail value for the low-cut gold bathing suit with a low back, and the weight is 5 hundred g.

at the weight of five hundred grams, the low cut, halter swimsuit made up of gold will sell at ten million Japanese Yen (JPY).

(Dreyer and Marcu, 2014)

Evaluation using paraphrases: the data

- HyTER: ~200 sentences, 500 translations each
- COCO: 5k images, 5 captions each

http://cocodataset.org/#explore?id=78026
(Lin et al., 2014)

a person is feeding a donut to the cat.

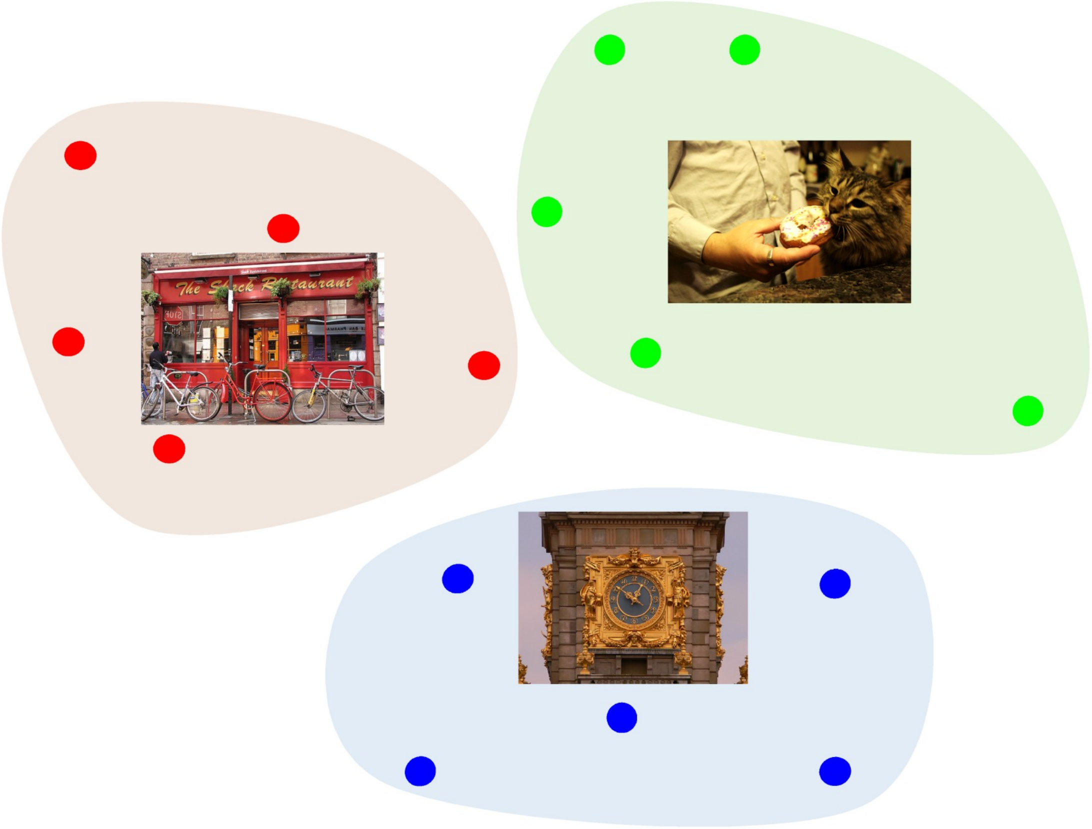
a cat being fed a donut by someone in a grey shirt.

a cat nibbles on a sprinkled donut that is being fed by the owner.

a grey cat biting into a frosted donuts

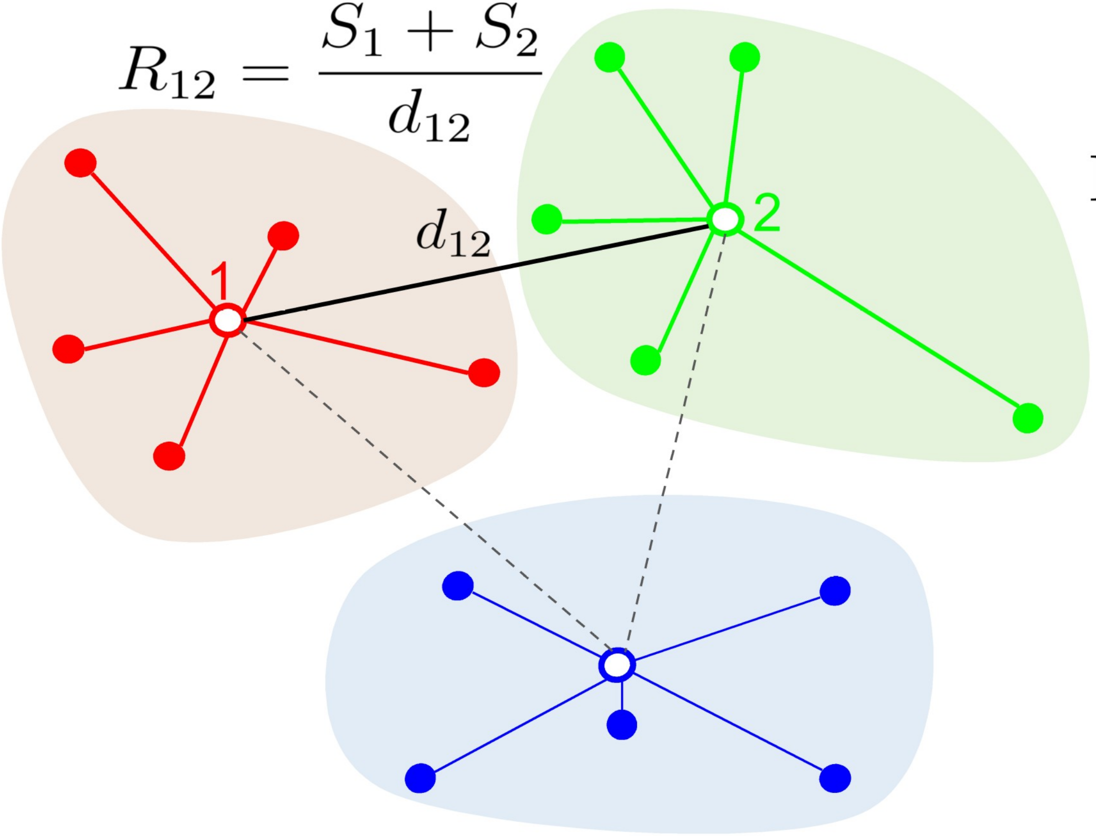
a cat is eating a donut from a person's hand.

Evaluation using paraphrases: the metrics



Cluster separation: Davies-Bouldin index

$$R_{12} = \frac{S_1 + S_2}{d_{12}}$$

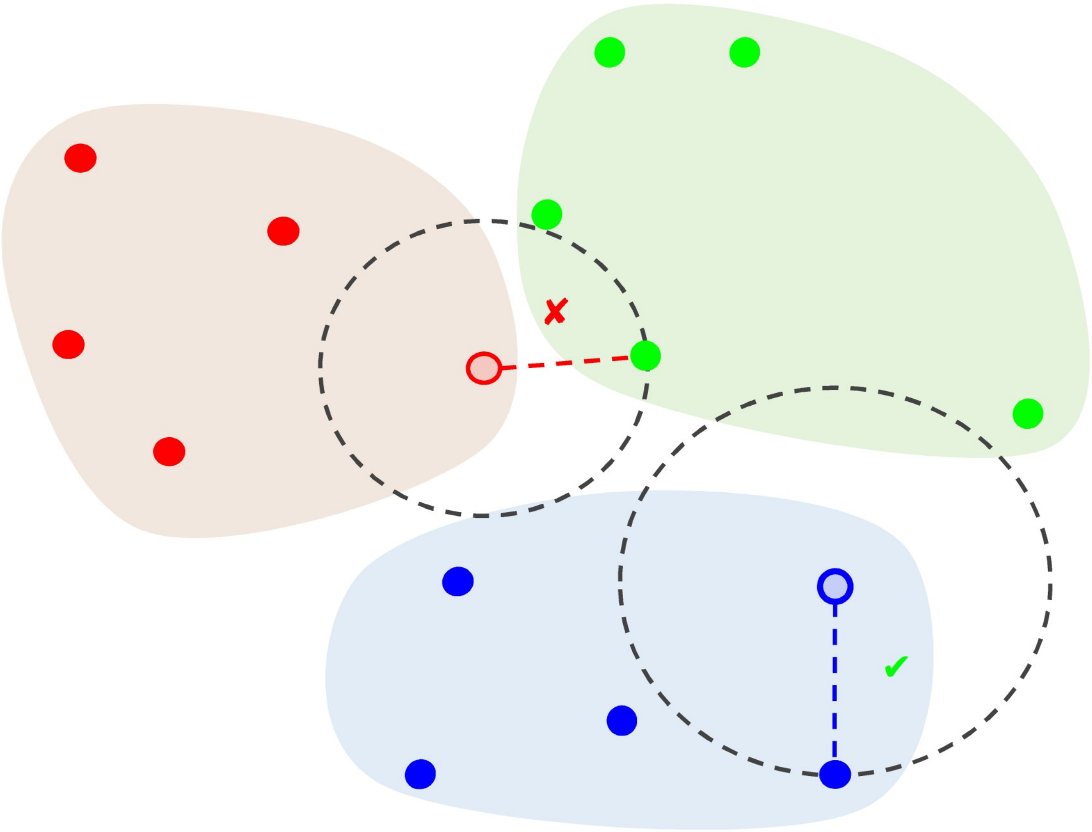


$$\mathrm{DB} = \frac{1}{N} \sum_{i=1}^{N} \max_{j \neq i} R_{ij}$$

For each cluster, find the **least well-separated** one

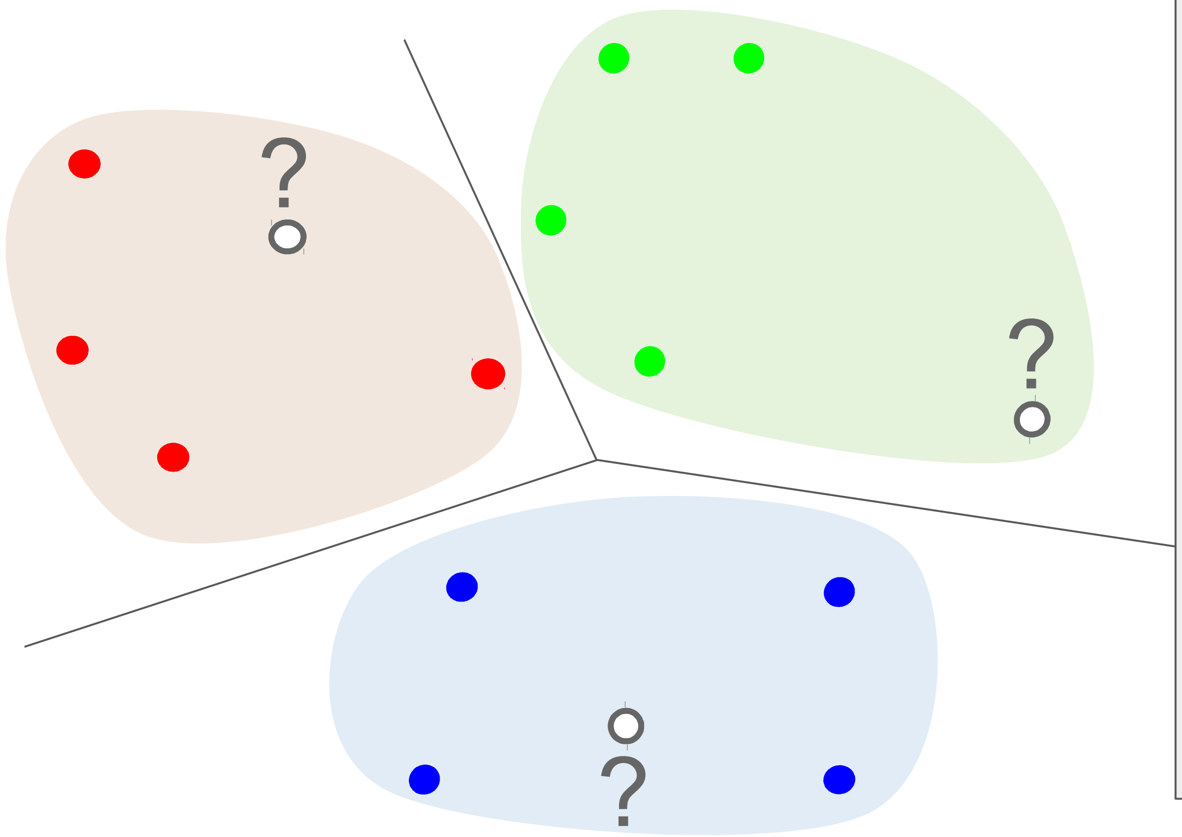
(Davies and Bouldin, 1979)

Paraphrase retrieval task (NN)



Retrieve the **nearest neighbor** and check whether it lies in the same cluster

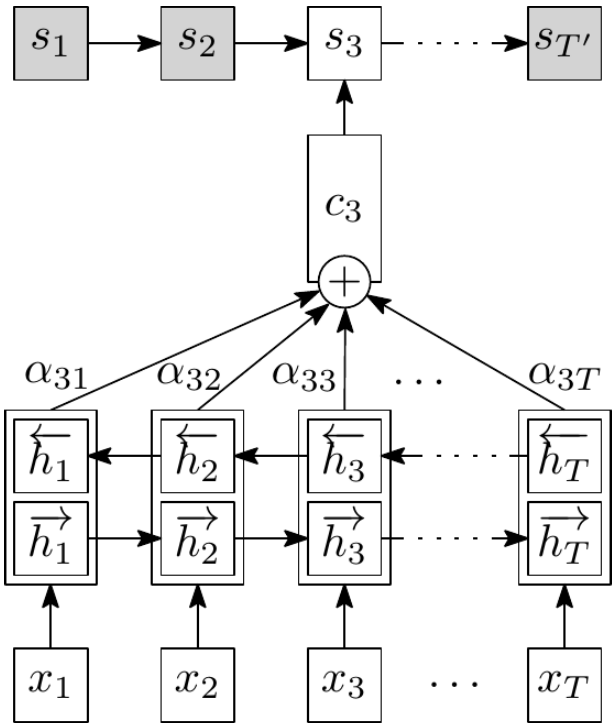
Classification task



1. Remove some points from the clusters.
2. Train an LDA classifier with the remaining points.
3. Classify the removed points back.

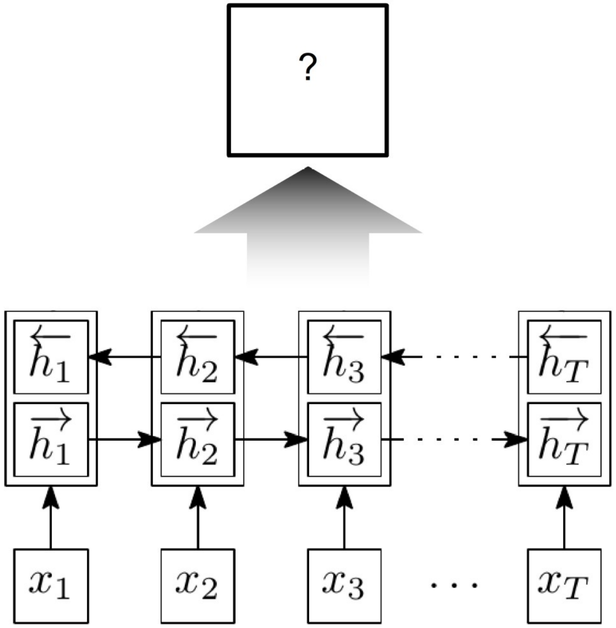
Sequence-to-sequence with attention

- Bahdanau et al. (2014)
- α_{ij}: weight of the j^{th} encoder state for the i^{th} decoder state
- no sentence embedding



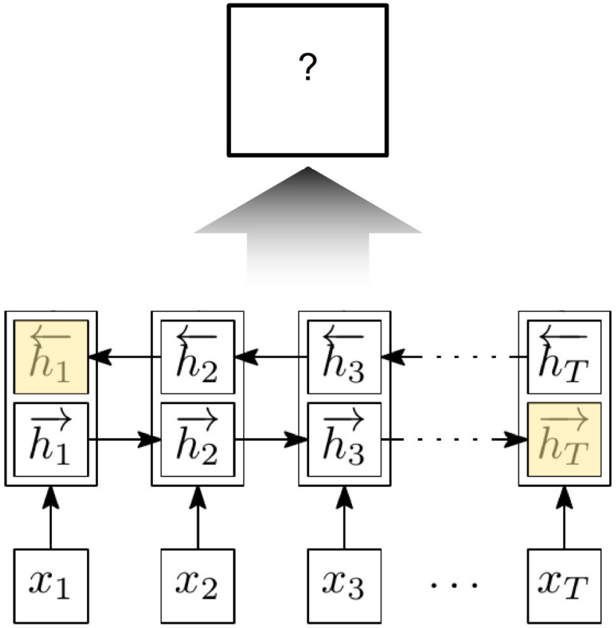
Ways of getting sentence embeddings

- final state
- max/average pooling
- inner attention



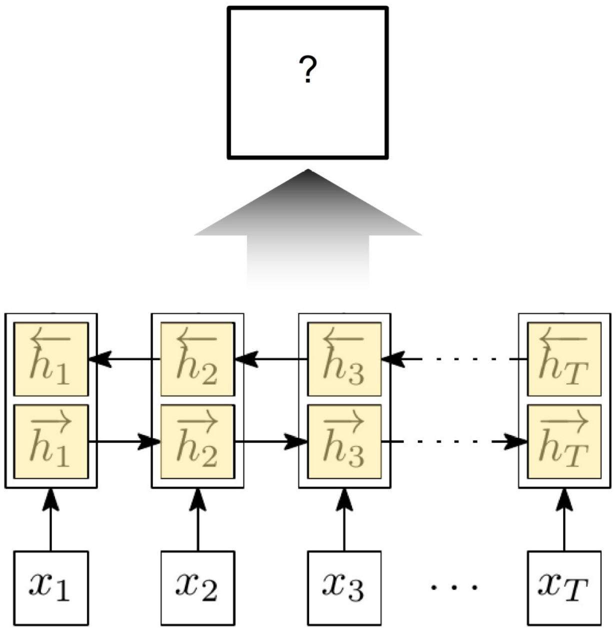
Ways of getting sentence embeddings

- final state
- max/average pooling
- inner attention



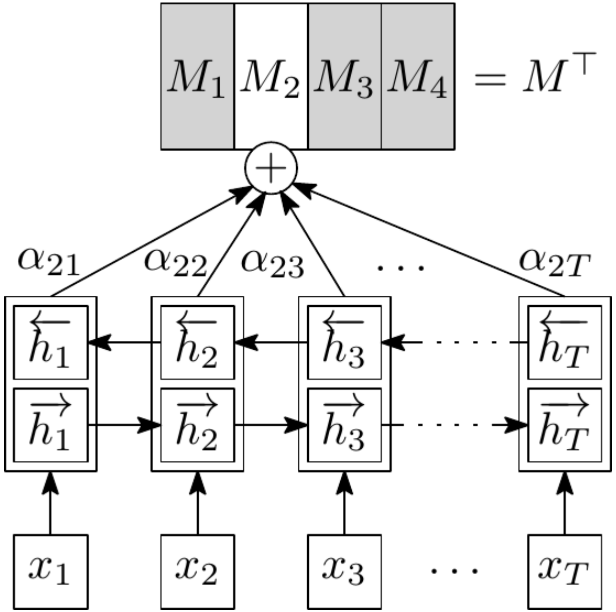
Ways of getting sentence embeddings

- final state
- max/average pooling
- inner attention

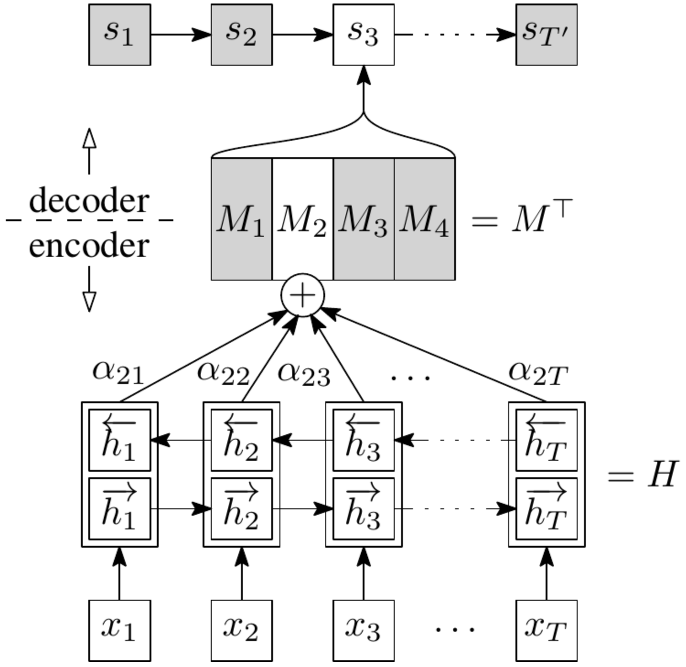


Multi-head inner attention

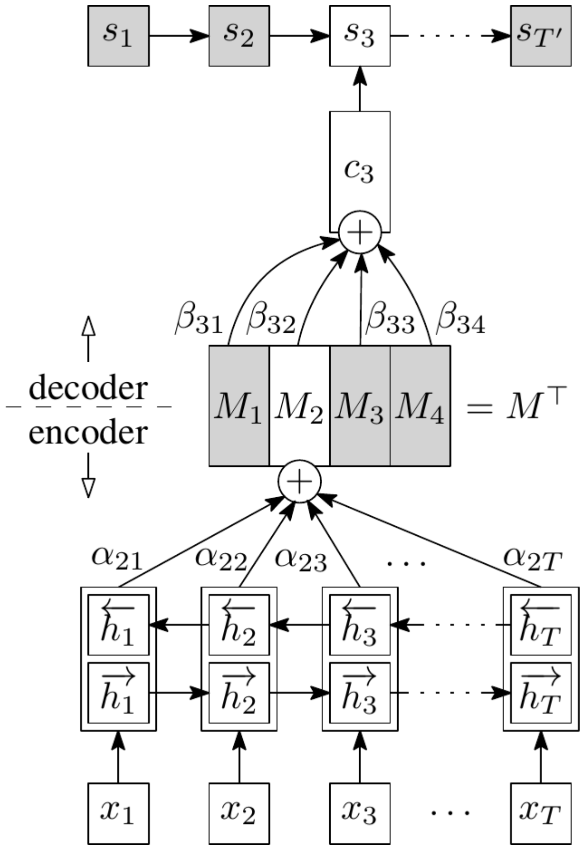
- Liu et al. (2016), Li et al. (2016), Lin et al. (2017)
- α_{ij}: weight of the j^{th} encoder state for the i^{th} column of $M^\top$
- concatenate columns of $M^\top$ → sentence embedding
- linear projection of columns to control embedding size



Proposed NMT architectures

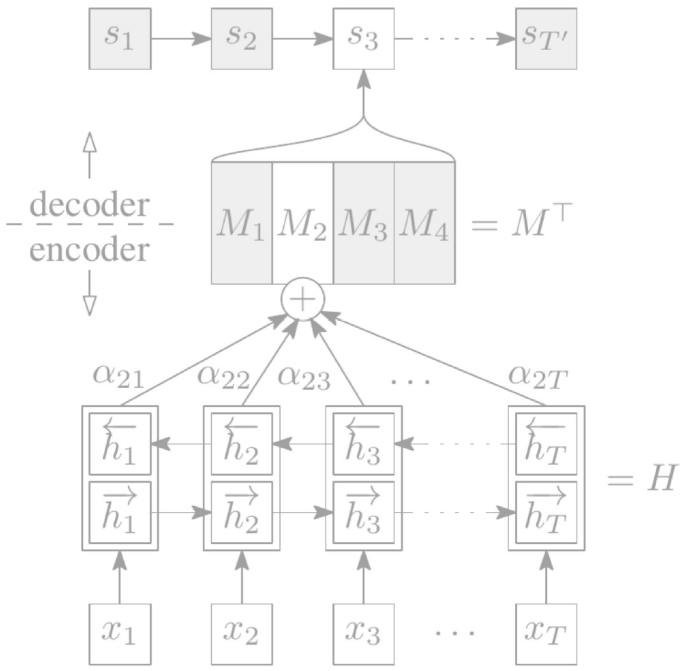


ATTN-CTX
decoder operates on entire embedding

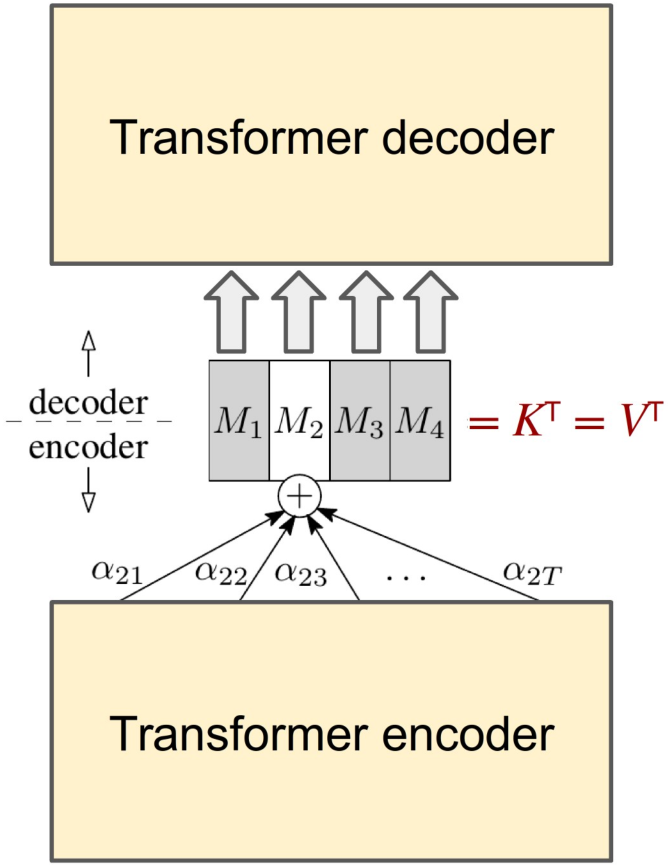


ATTN-ATTN (*compound att.)*
decoder „selects“ components of embedding

Proposed NMT architectures



ATTN-CTX
decoder operates on entire embedding



TRF-ATTN-ATTN
Transformer (Vaswani et al., 2017) with inner attention

Evaluated NMT models

- model architectures:
  - **FINAL**, **FINAL-CTX**: no attention
  - **AVGPOOL**, **MAXPOOL**: pooling instead of attention
  - **ATTN-CTX**: inner attention, constant context vector
  - **ATTN-ATTN**: inner attention, decoder attention
  - **TRF-ATTN-ATTN**: Transformer with inner attention
- translation from English (to Czech or German), evaluating embeddings of English (source) sentences
  - en→cs: CzEng 1.7 (Bojar et al., 2016)
  - en→de: Multi30K (Elliott et al., 2016; Helcl and Libovický, 2017)

Sample Results – translation quality en→cs

	Model	Heads	BLEU	Manual (> other)	Manual (≥ other)
„Bahdanau“	ATTN	—	**22.2**	**50.9**	**93.8**
compound attention	ATTN-ATTN	8	18.4	42.5	88.6
	ATTN-ATTN	4	17.1	—	—
inner attention + „Cho“	ATTN-CTX	4	16.1	31.7	77.9
„Cho“	FINAL-CTX	—	15.5	—	—
	ATTN-ATTN	1	14.8	27.3	71.7
„Sutskever“	FINAL	—	10.8	—	—

Selected models trained for translation from English to Czech. The embedding size is 1000 (except ATTN).

Sample Results – translation quality en→cs

	Model	Heads	BLEU	Manual (> other)	Manual (≥ other)
„Bahdanau“	ATTN	—	**22.2**	**50.9**	**93.8**
compound attention	ATTN-ATTN	8	<u>18.4</u>	<u>42.5</u>	<u>88.6</u>
	ATTN-ATTN	4	17.1	—	—
inner attention + „Cho“	ATTN-CTX	4	16.1	31.7	77.9
„Cho“	FINAL-CTX	—	15.5	—	—
	ATTN-ATTN	1	14.8	27.3	71.7
„Sutskever“	FINAL	—	10.8	—	—

BLEU is consistent with human evaluation.

Selected models trained for translation from English to Czech. The embedding size is 1000 (except ATTN).

Sample Results – translation quality en→cs

	Model	Heads	BLEU	Manual (> other)	Manual (≥ other)
„Bahdanau“	ATTN	—	**22.2**	**50.9**	**93.8**
compound attention	ATTN-ATTN	8	<u>18.4</u>	<u>42.5</u>	<u>88.6</u>
	ATTN-ATTN	4	17.1	—	—
inner attention + „Cho“	ATTN-CTX	4	16.1	31.7	77.9
„Cho“	FINAL-CTX	—	15.5	—	—
	ATTN-ATTN	1	14.8	27.3	71.7
„Sutskever“	FINAL	—	10.8	—	—

Attention in the encoder helps translation quality.

Selected models trained for translation from English to Czech. The embedding size is 1000 (except ATTN).

Sample Results – translation quality en→cs

	Model	Heads	BLEU	Manual (> other)	Manual (≥ other)
„Bahdanau“	ATTN	—	**22.2**	**50.9**	**93.8**
compound attention	ATTN-ATTN	8	<u>18.4</u>	<u>42.5</u>	<u>88.6</u>
	ATTN-ATTN	4	17.1	—	—
inner attention + „Cho“	ATTN-CTX	4	16.1	31.7	77.9
„Cho“	FINAL-CTX	—	15.5	—	—
	ATTN-ATTN	1	14.8	27.3	71.7
„Sutskever“	FINAL	—	10.8	—	—

More attention heads
→ better translation quality.

Selected models trained for translation from English to Czech. The embedding size is 1000 (except ATTN).

Sample Results – representation eval. en→cs

Model	Size	Heads	**SentEval** AvgAcc	**SentEval** AvgSim	**Paraphrases** class. accuracy (COCO)
InferSent	4096	—	**81.7**	**0.70**	31.58
GloVe bag-of-words	300	—	75.8	0.59	**34.28**
FINAL-CTX (“Cho“)	1000	—	**74.4**	**0.60**	**23.20**
ATTN-ATTN	1000	1	73.4	0.54	21.54
ATTN-CTX	1000	4	72.2	0.45	14.60
ATTN-ATTN	1000	4	70.8	0.39	10.84
ATTN-ATTN	1000	8	70.0	0.36	10.24

Selected models trained for translation from English to Czech. InferSent and GloVe-BOW are trained on monolingual (English) data.

Sample Results – representation eval. en→cs

Model	Size	Heads	**SentEval** AvgAcc	**SentEval** AvgSim	**Paraphrases** class. accuracy (COCO)
InferSent	4096	—	**81.7**	**0.70**	31.58
GloVe bag-of-words	300	—	75.8	0.59	**34.28**
FINAL-CTX (“Cho“)	1000	—	**74.4**	**0.60**	**23.20**
ATTN-ATTN	1000	1	73.4	0.54	21.54
ATTN-CTX	1000	4	72.2	0.45	14.60
ATTN-ATTN	1000	4	70.8	0.39	10.84
ATTN-ATTN	1000	8	70.0	0.36	10.24

Baselines are hard to beat.

Selected models trained for translation from English to Czech. InferSent and GloVe-BOW are trained on monolingual (English) data.

Sample Results – representation eval. en→cs

Model	Size	Heads	**SentEval** AvgAcc	**SentEval** AvgSim	**Paraphrases** class. accuracy (COCO)
InferSent	4096	—	**81.7**	**0.70**	31.58
GloVe bag-of-words	300	—	75.8	0.59	**34.28**
FINAL-CTX (“Cho“)	1000	—	**74.4**	**0.60**	**23.20**
ATTN-ATTN	1000	1	73.4	0.54	21.54
ATTN-CTX	1000	4	72.2	0.45	14.60
ATTN-ATTN	1000	4	70.8	0.39	10.84
ATTN-ATTN	1000	8	70.0	0.36	10.24

Attention harms the performance.

Selected models trained for translation from English to Czech. InferSent and GloVe-BOW are trained on monolingual (English) data.

Sample Results – representation eval. en→cs

Model	Size	Heads	**SentEval** AvgAcc	**SentEval** AvgSim	**Paraphrases** class. accuracy (COCO)
InferSent	4096	—	**81.7**	**0.70**	31.58
GloVe bag-of-words	300	—	75.8	0.59	**34.28**
FINAL-CTX (“Cho“)	1000	—	**74.4**	**0.60**	**23.20**
ATTN-ATTN	1000	1	73.4	0.54	21.54
ATTN-CTX	1000	4	72.2	0.45	14.60
ATTN-ATTN	1000	4	70.8	0.39	10.84
ATTN-ATTN	1000	8	70.0	0.36	10.24

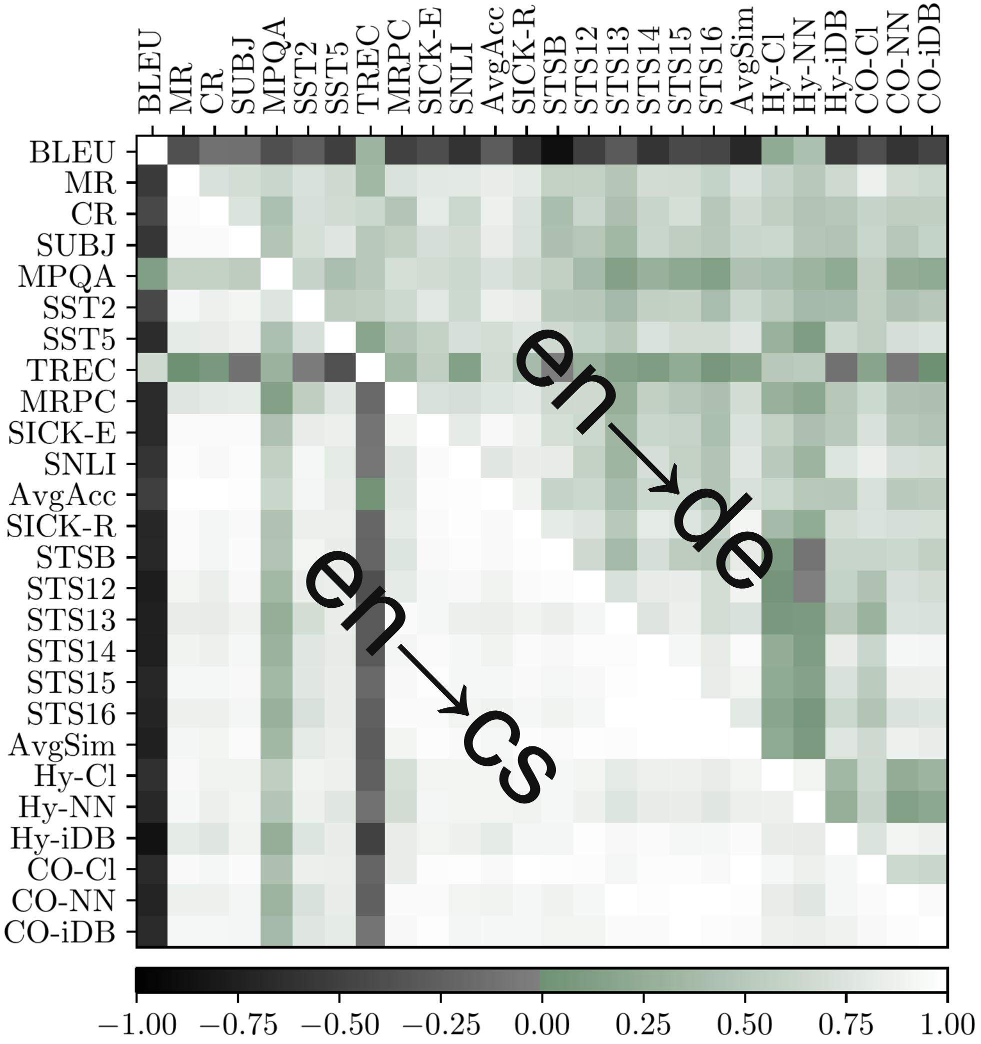
More heads → worse results.

Selected models trained for translation from English to Czech. InferSent and GloVe-BOW are trained on monolingual (English) data.

Full Results – correlations

BLEU vs. other metrics:
−0.57 ± 0.31 (en→**cs**)
−0.36 ± 0.29 (en→**de**)

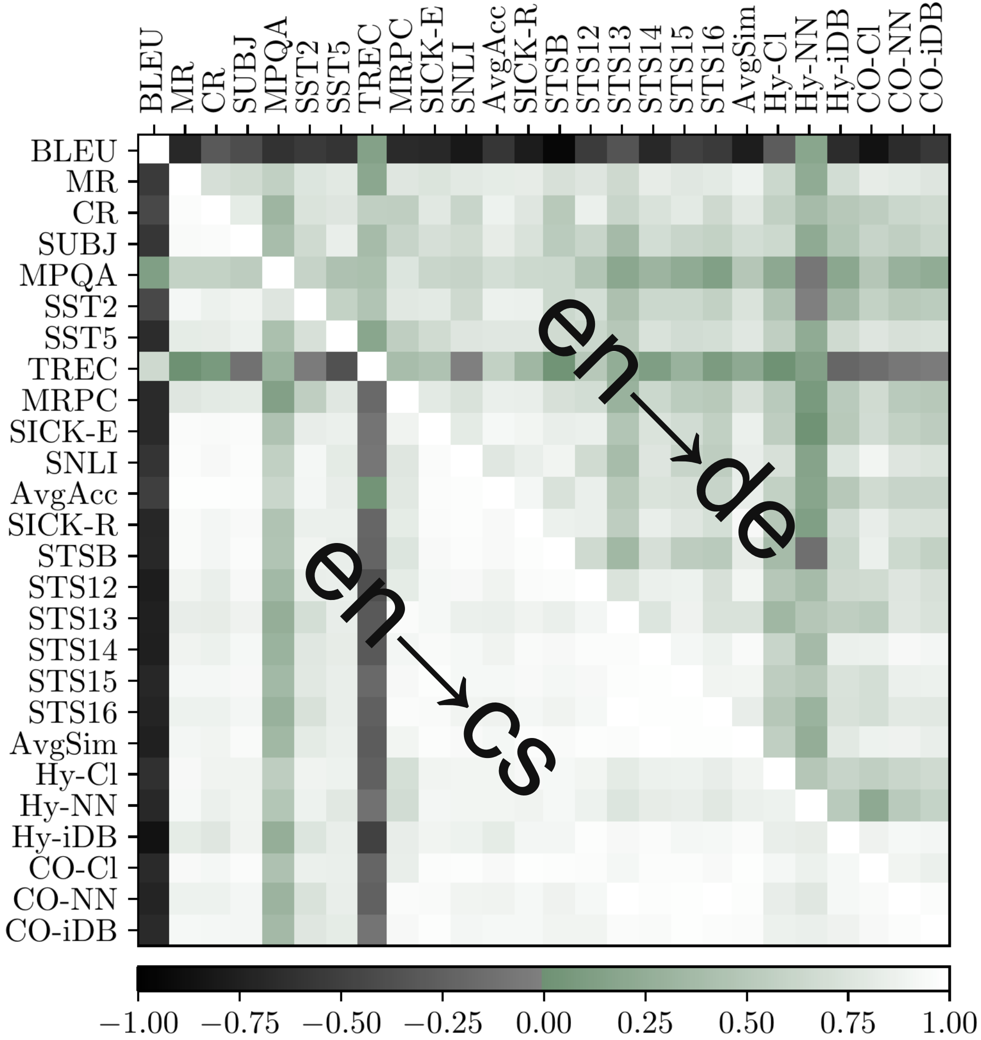
Pairwise average (except BLEU):
0.78 ± 0.32 (en→**cs**)
0.57 ± 0.23 (en→**de**)



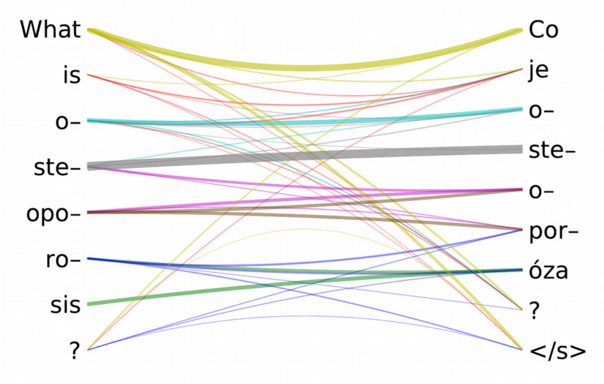
Full Results – correlations excluding Transformer

BLEU vs. other metrics:
−0.57 ± 0.31 (en→**cs**)
−0.54 ± 0.27 (en→**de**)

Pairwise average (except BLEU):
0.78 ± 0.32 (en→**cs**)
0.62 ± 0.23 (en→**de**)

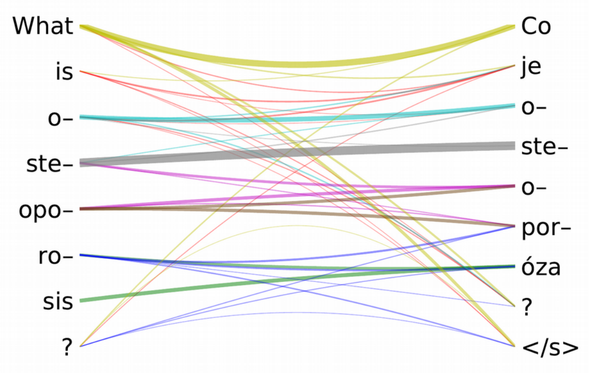


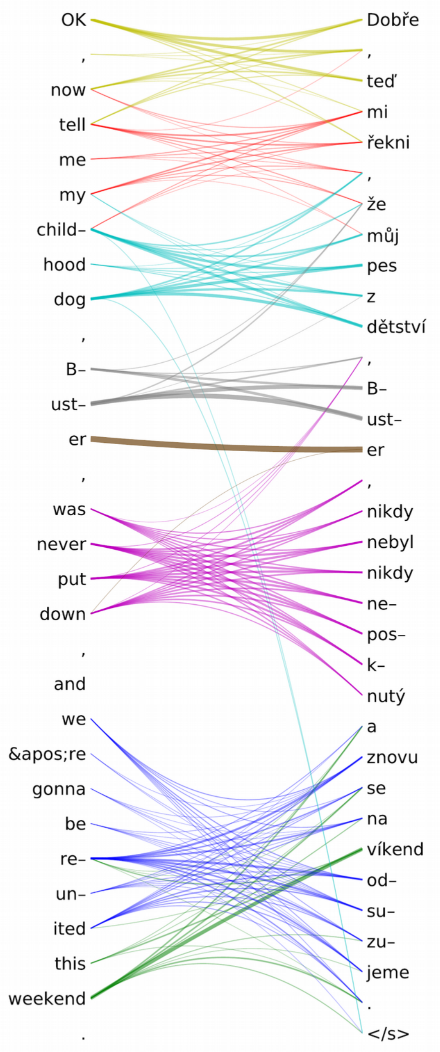
Compound attention interpretation

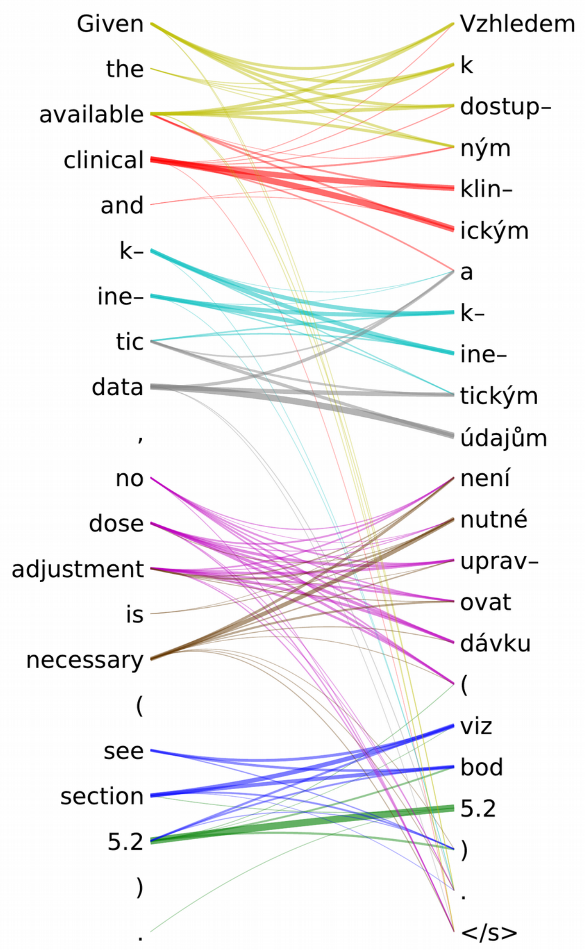


ATTN-ATTN en-cs model with 8 heads

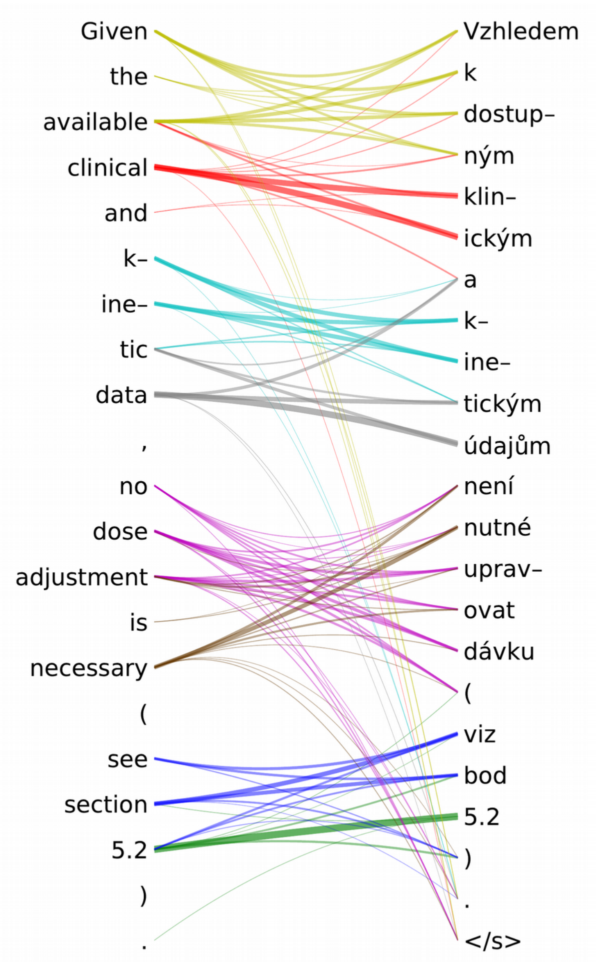
Compound attention interpretation

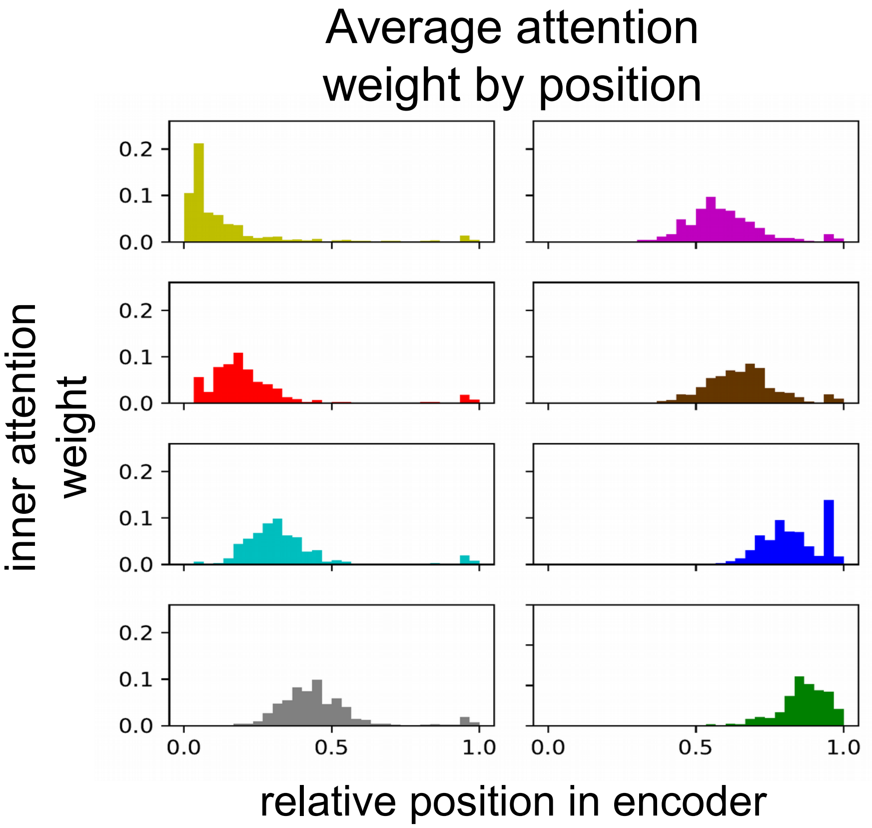


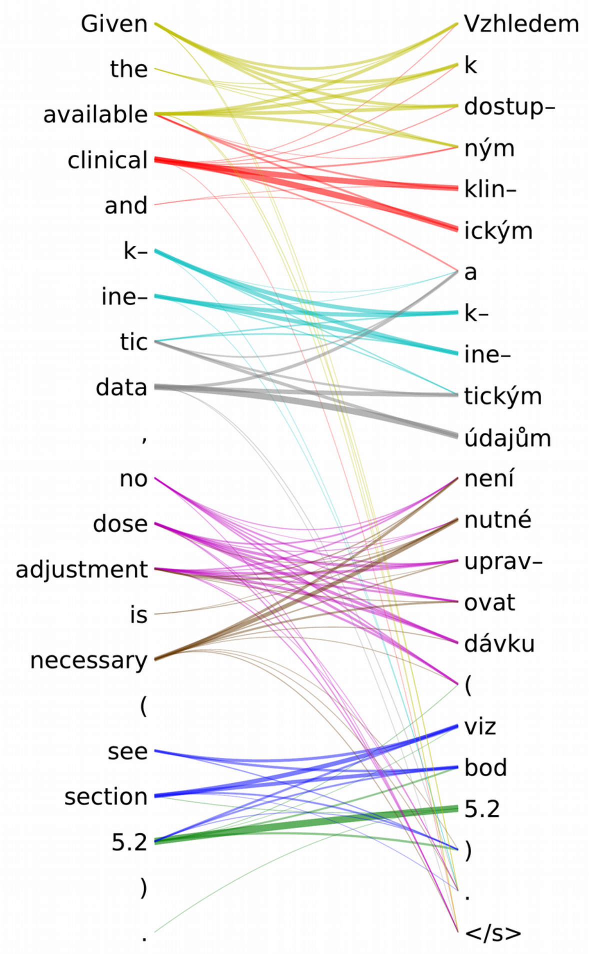




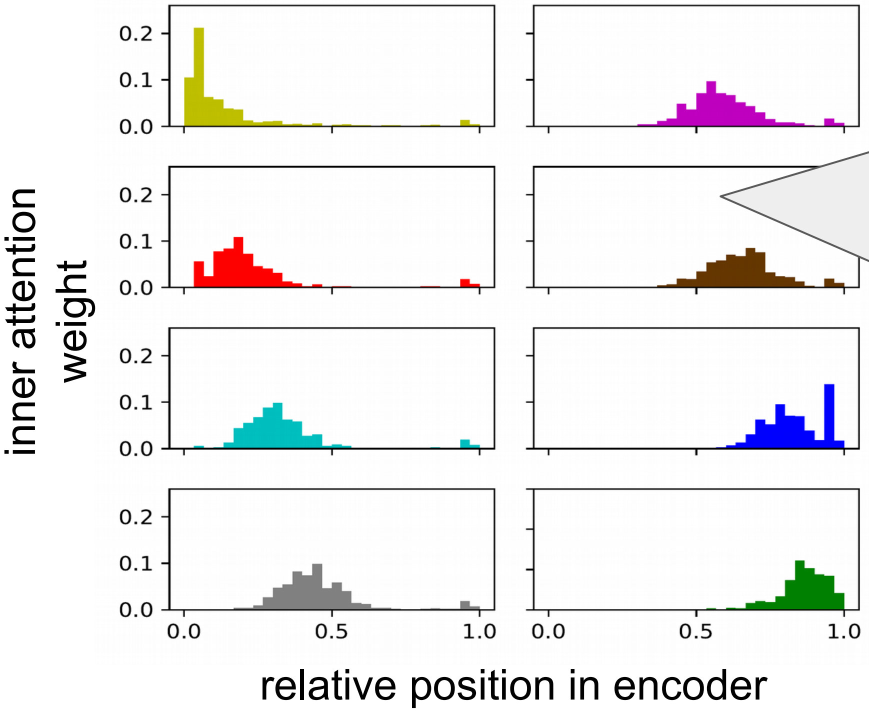
ATTN-ATTN en-cs model with 8 heads

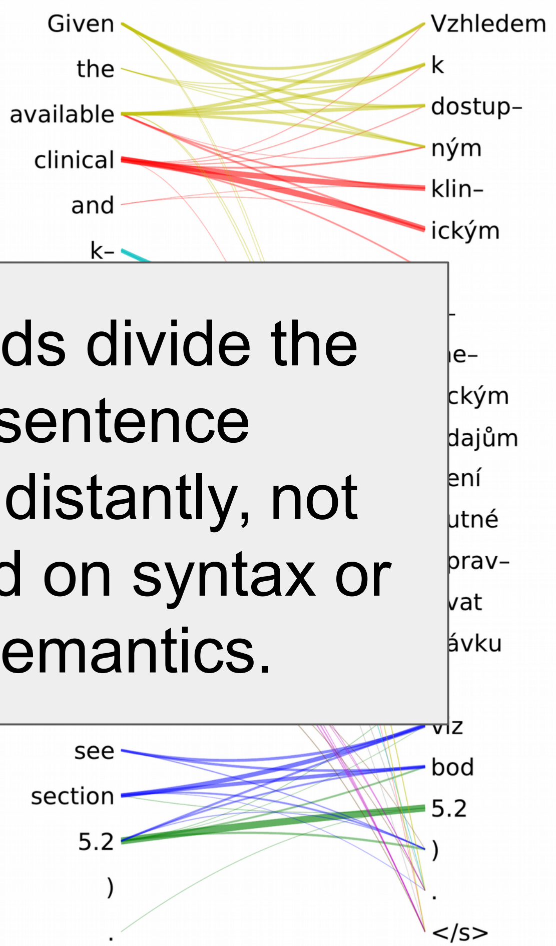

Given
the
available
clinical
and
k-
ine-
tic
data
,
no
dose
adjustment
is
necessary
(
see
section
5.2
)
.
Vzhledem
k
dostup-
ným
klin-
ickým
a
k-
ine-
tickým
údajům
není
nutné
uprav-
ovat
dávku
(
viz
bod
5.2
)
.
</s>


Average attention
weight by position
inner attention
weight
0.2
0.1
0.0
0.0
0.5
1.0
relative position in encoder


Given
the
available
clinical
and
k-
ine-
tic
data
,
no
dose
adjustment
is
necessary
(
see
section
5.2
)
.
Vzhledem
k
dostup-
ným
klin-
ickým
a
k-
ine-
tickým
údajům
není
nutné
uprav-
ovat
dávku
(
viz
bod
5.2
)
.
</s>

Average attention weight by position





Heads divide the sentence equidistantly, not based on syntax or semantics.

Summary

Summary

- Proposed NMT architecture combining the benefit of attention and one $&!#* vector representing the whole sentence.

Summary

- Proposed NMT architecture combining the benefit of attention and one $&!#* vector representing the whole sentence.
- Evaluated the obtained sentence embeddings using a wide range of “semantic” tasks.

Summary

- Proposed NMT architecture combining the benefit of attention and one $&!#* vector representing the whole sentence.
- Evaluated the obtained sentence embeddings using a wide range of “semantic” tasks.
- The better the translation, the worse performance in “meaning” representation.

Summary

- Proposed NMT architecture combining the benefit of attention and one $&!#* vector representing the whole sentence.
- Evaluated the obtained sentence embeddings using a wide range of “semantic” tasks.
- The better the translation, the worse performance in “meaning” representation.
- Heads divide sentence equidistantly, not logically.

Summary

- Proposed NMT architecture combining the benefit of attention and one $&!#* vector representing the whole sentence.
- Evaluated the obtained sentence embeddings using a wide range of “semantic” tasks.
- The better the translation, the worse performance in
-

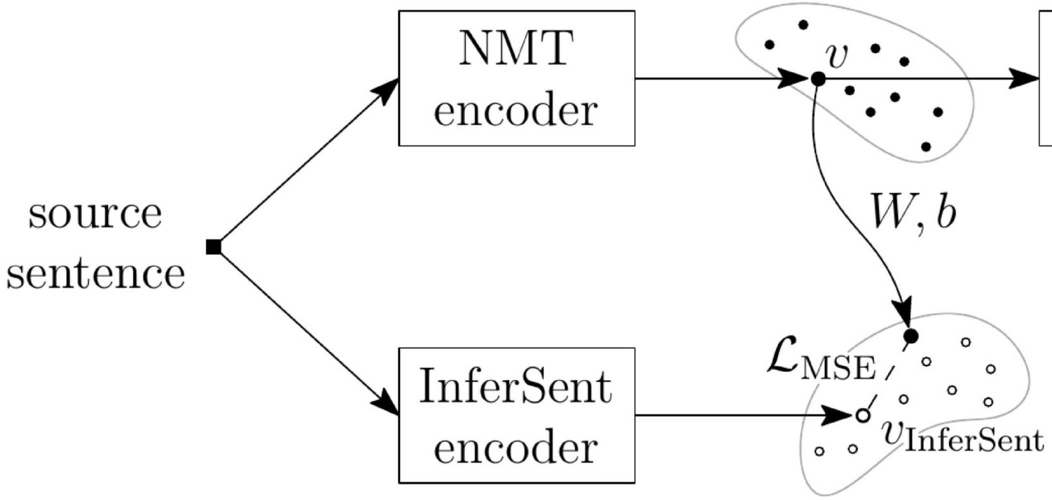
Join our

JNLE Special Issue on Sentence Representations:

http://ufal.mff.cuni.cz/jnle-on-sentence-representation

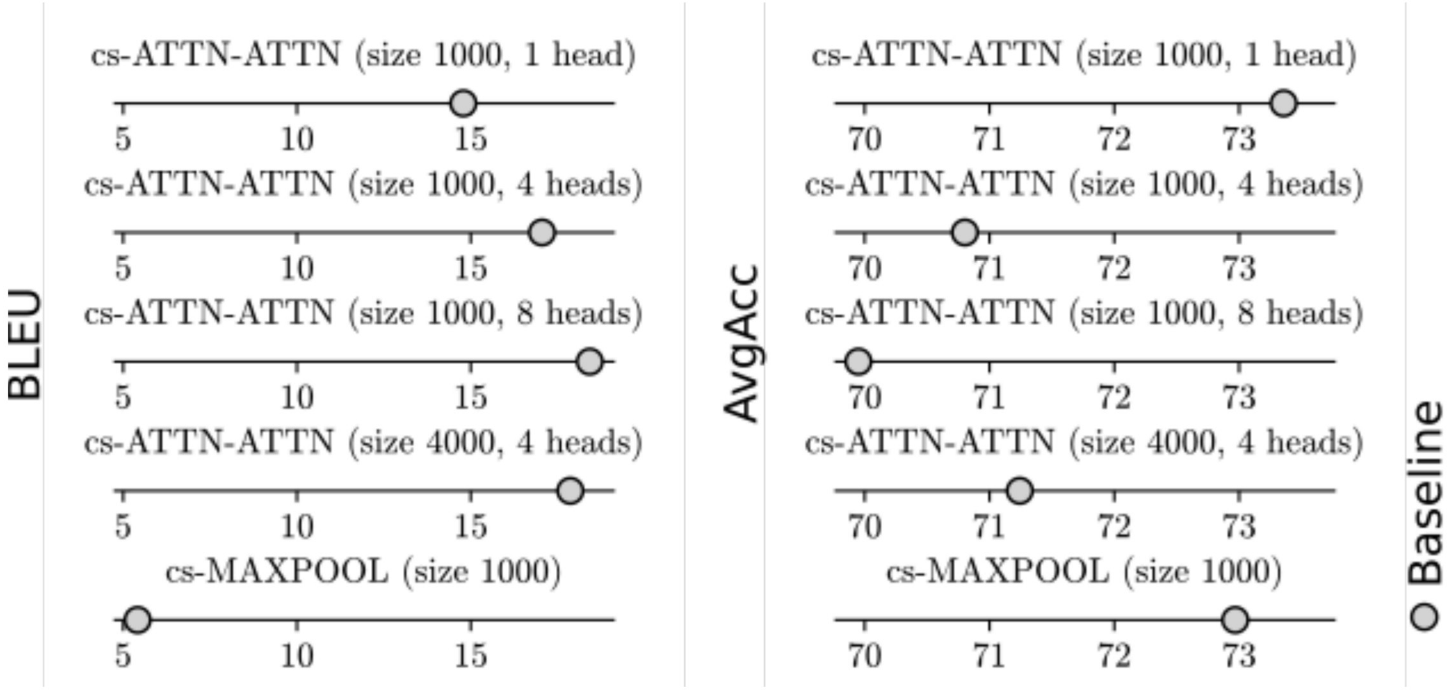
InferSent multi-task training (in OC’s thesis only)

- Idea: produce better representations by jointly training NMT with other tasks
- Proxy: predict InferSent embeddings as the auxiliary task

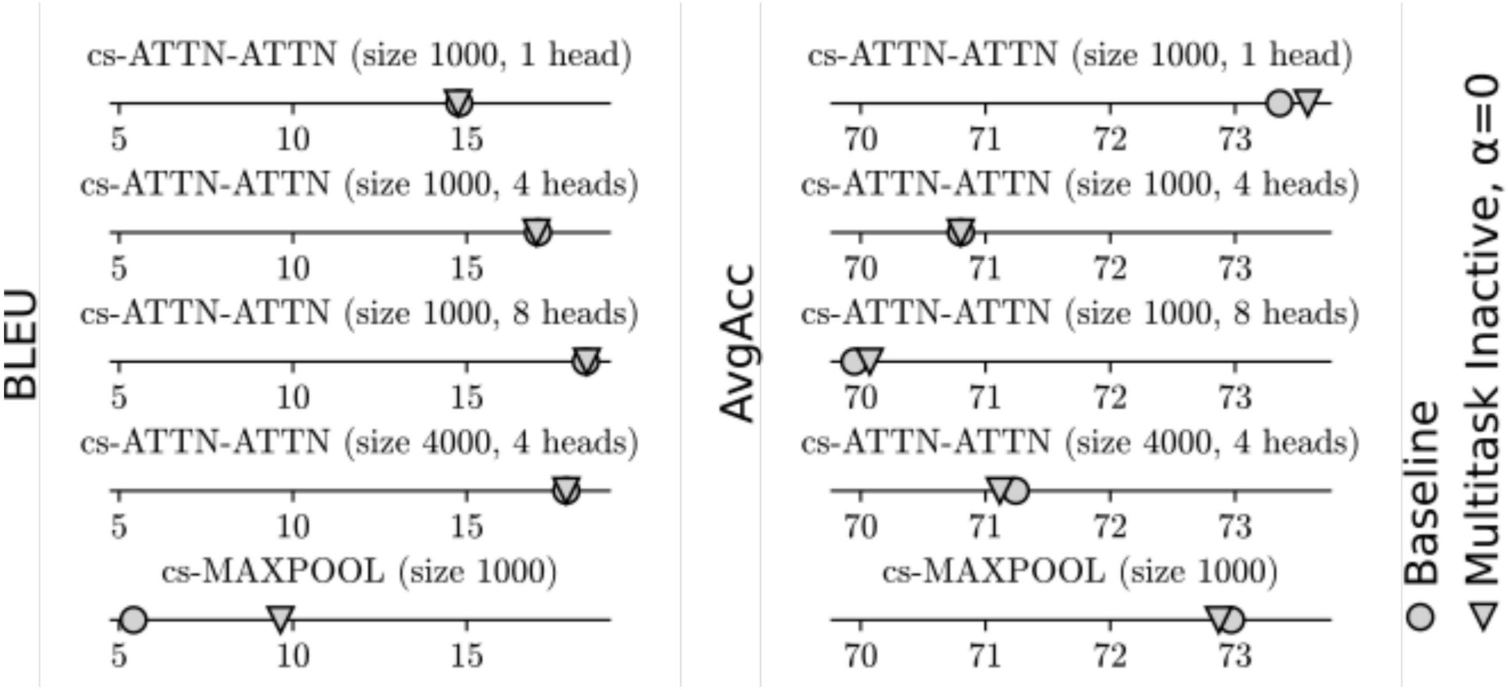


$$\mathcal{L} = \mathcal{L}_{\mathrm{MT}} + \alpha \mathcal{L}_{\mathrm{MSE}}$$

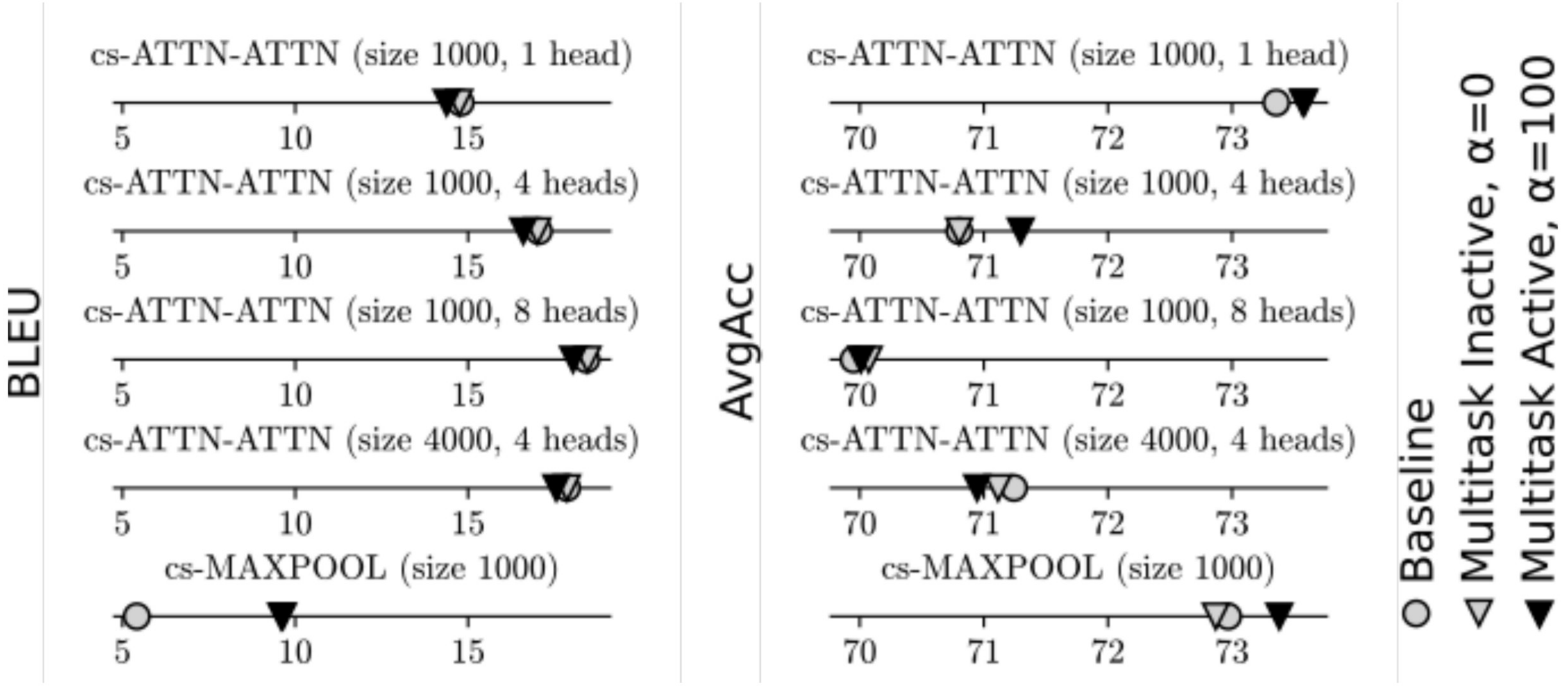
Multi-task training results en→cs



Multi-task training results en→cs

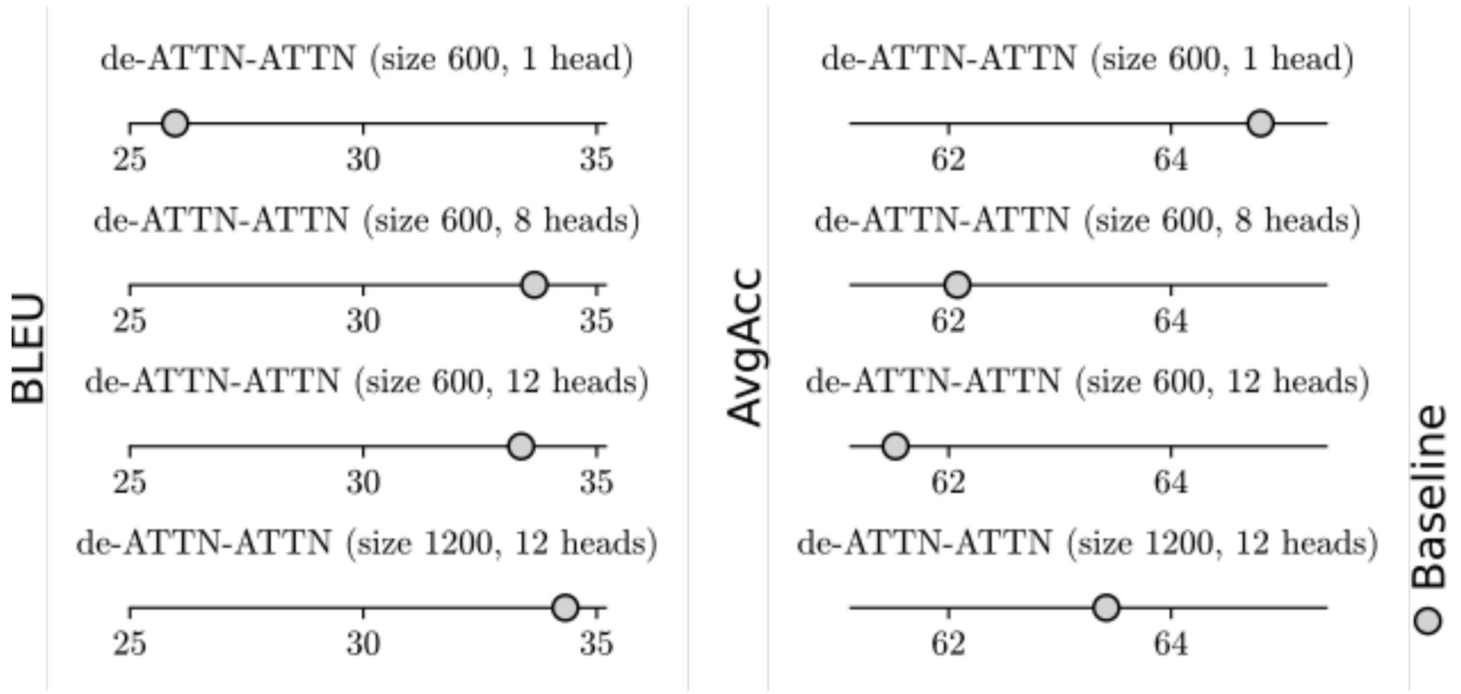


Multi-task training results en→cs

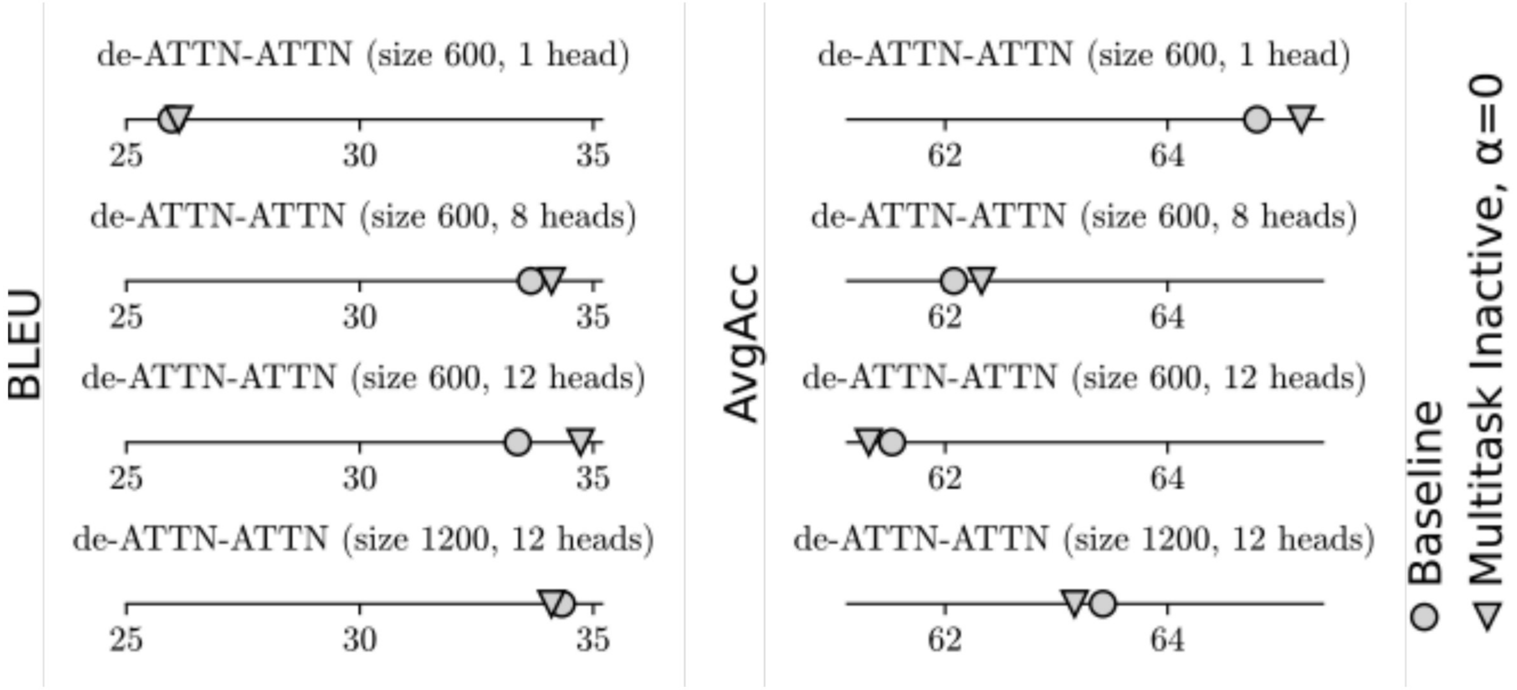


→ Small loss in BLEU (exc. MAXPOOL), sometimes gain in AvgAcc (exc. 4000, 4h)

Multi-task training results en→de

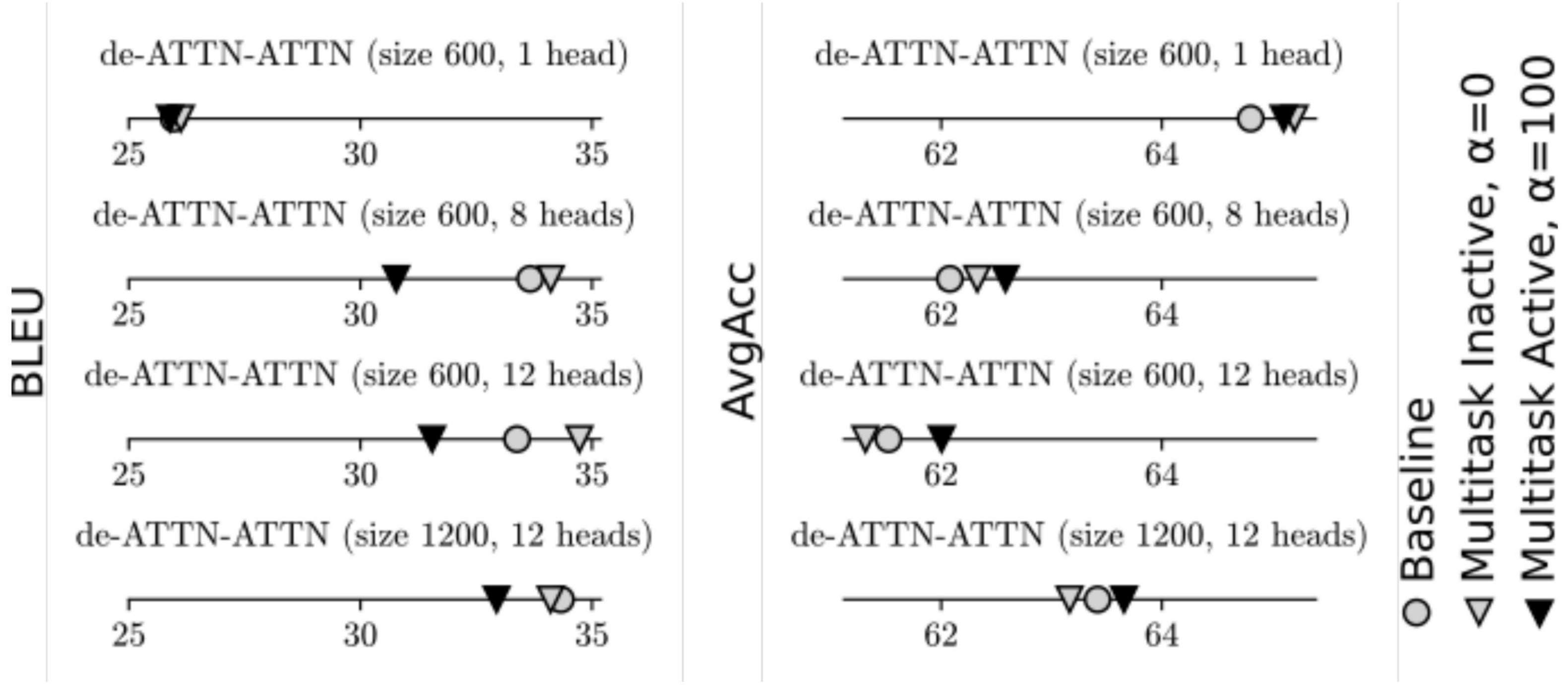


Multi-task training results en→de



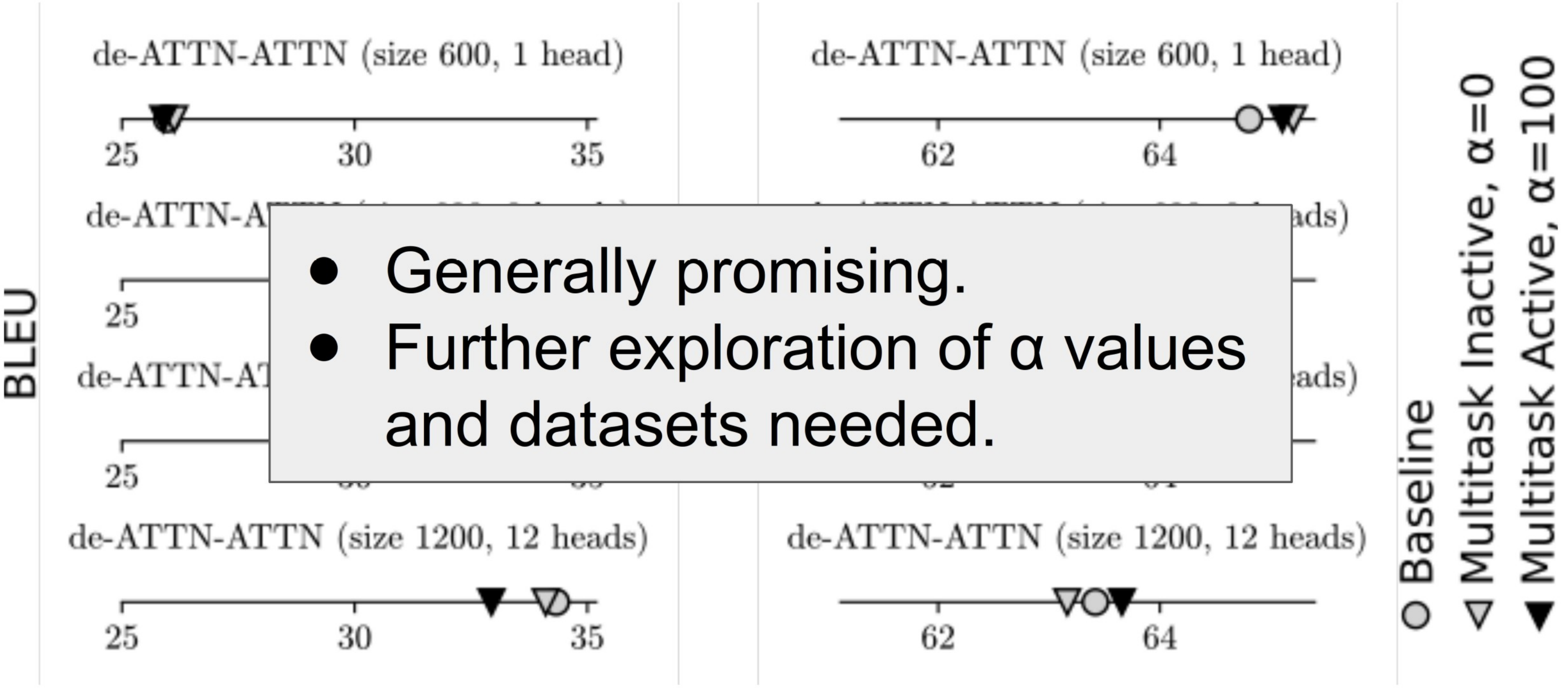
→ en-de results less stable (much smaller vocabulary).

Multi-task training results en→de



→ Big loss in BLEU (exc. 600, 1h), small gain in AvgAcc (exc. 600, 1h)

Multi-task training results en→de



- Generally promising.
- Further exploration of α values and datasets needed.

→ Big loss in BLEU (exc. 600, 1h), small gain in AvgAcc (exc. 600, 1h)

Bibliography

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. *Neural machine translation by jointly learning to align and translate.* In ICLR.

Ondřej Bojar et al. 2016. *CzEng 1.6: Enlarged Czech-English parallel corpus with processing tools dockered.* In Text, Speech, and Dialogue (TSD), number 9924 in LNAI, pages 231–238.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. *Learning phrase representations using rnn encoder-decoder for statistical machine translation.* In EMNLP.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. *Supervised learning of universal sentence representations from natural language inference data.* In EMNLP.

David L. Davies and Donald W. Bouldin. *A cluster separation measure.* IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1:224–227, 1979.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. *Multi30k: Multilingual English-German image descriptions.* CoRR, abs/1605.00459.

Jindřich Helcl and Jindřich Libovický. 2017. *CUNI System for the WMT17 Multimodal Traslation Task.*

Bibliography

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. *Skip-thought vectors.* In NIPS Vol. 2, NIPS'15.

Markus Dreyer and Daniel Marcu. 2014. *HyTER networks of selected OpenMT08/09 sentences.* Linguistic Data Consortium. LDC2014T09.

Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. 2016. *Dataset and neural recurrent sequence labeling model for open-domain factoid question answering.* CoRR, abs/1607.06275.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. 2014. *Microsoft COCO: common objects in context.* CoRR, abs/1405.0312.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. *A structured self-attentive sentence embedding.* CoRR, abs/1703.03130.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. *Learning natural language inference using bidirectional LSTM model and inner-attention.* CoRR, abs/1605.09090.

Bibliography

Holger Schwenk and Matthijs Douze. 2017. *Learning joint multilingual sentence representations with neural machine translation.* CoRR, abs/1704.04154.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. *Sequence to sequence learning with neural networks.* In NIPS.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need.* In NIPS.