Are BLEU and Meaning Representation in Opposition?

Ondřej Cífka
Ondřej Bojar
Motivation

- Good translation preserves the meaning of the sentence.
- Neural MT learns to represent the sentence.
  - Is the representation “meaningful” in some sense?
Motivation: The observed genome representation “meat graph”
Motivation:

- Serves as a bridge between representation and semantics.
- For a given sentence representation, “metaphorically”.

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

SentEval Classification Tasks

an ambitious and moving but bleak film.
and that makes all the difference.
rarely, a movie is more than a movie.
the movie is well done, but slow.
the pianist is polanski 's best film.
Evaluation through classification

SentEval Classification Tasks

an ambitious and moving but bleak film. and that makes all the difference. rarely, a movie is more than a movie. the movie is well done, but slow. the pianist is polanski’s best film.
Evaluation through classification

**SentEval Classification Tasks**

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>an ambitious and moving but bleak film.</td>
<td>1</td>
</tr>
<tr>
<td>and that makes all the difference.</td>
<td>0</td>
</tr>
<tr>
<td>rarely, a movie is more than a movie.</td>
<td>1</td>
</tr>
<tr>
<td>the movie is well done, but slow.</td>
<td>0</td>
</tr>
<tr>
<td>the pianist is polanski's best film.</td>
<td>2</td>
</tr>
</tbody>
</table>

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

**SentEval Classification Tasks**

- A square full of people and life.
- The square is busy.
- The couple is at a restaurant.
- A cute couple at a club
- A white dog bounding through snow

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

**SentEval Classification Tasks**

A square full of people and life.
The square is busy.
The couple is at a restaurant.
A cute couple at a club
A white dog bounding through snow

- 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

<table>
<thead>
<tr>
<th>Sentence 1</th>
<th>Sentence 2</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>I think it probably depends on your money.</td>
<td>It depends on your country.</td>
<td>0</td>
</tr>
<tr>
<td>Yes, you should mention your experience.</td>
<td>Yes, you should make a resume</td>
<td>2</td>
</tr>
<tr>
<td>Hope this is what you are looking for.</td>
<td>Is this the kind of thing you're looking for?</td>
<td>4</td>
</tr>
</tbody>
</table>

- 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

![Image of a person feeding a donut to a cat](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

\[ R_{12} = \frac{S_1 + S_2}{d_{12}} \]

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

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.
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)
- $\alpha_{ij}$: weight of the $j^{\text{th}}$ encoder state for the $i^{\text{th}}$ 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)
- $\alpha_{ij}$: 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Heads</th>
<th>BLEU</th>
<th>Manual (&gt; other)</th>
<th>Manual (≥ other)</th>
</tr>
</thead>
<tbody>
<tr>
<td>„Bahdanau“ ATTN</td>
<td>—</td>
<td>22.2</td>
<td>50.9</td>
<td>93.8</td>
</tr>
<tr>
<td>compound attention</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATTN-ATTN 8</td>
<td></td>
<td>18.4</td>
<td>42.5</td>
<td>88.6</td>
</tr>
<tr>
<td>ATTN-ATTN 4</td>
<td></td>
<td>17.1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>inner attention + „Cho“</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATTN-CTX 4</td>
<td></td>
<td>16.1</td>
<td>31.7</td>
<td>77.9</td>
</tr>
<tr>
<td>„Cho“ FINAL-CTX</td>
<td>—</td>
<td>15.5</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>„Sutskever“ ATTN-ATTN 1</td>
<td></td>
<td>14.8</td>
<td>27.3</td>
<td>71.7</td>
</tr>
<tr>
<td>FINAL</td>
<td>—</td>
<td>10.8</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Selected models trained for translation from English to Czech. The embedding size is 1000 (except ATTN).
Sample Results – translation quality en→cs

<table>
<thead>
<tr>
<th>Model</th>
<th>Heads</th>
<th>BLEU</th>
<th>Manual (&gt; other)</th>
<th>Manual (≥ other)</th>
</tr>
</thead>
<tbody>
<tr>
<td>„Bahdanau“</td>
<td>—</td>
<td>22.2</td>
<td>50.9</td>
<td>93.8</td>
</tr>
<tr>
<td>compound attention</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATTN-ATTN</td>
<td>8</td>
<td>18.4</td>
<td>42.5</td>
<td>88.6</td>
</tr>
<tr>
<td>ATTN-ATTN</td>
<td>4</td>
<td>17.1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>inner attention + „Cho“</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATTN-CTX</td>
<td>4</td>
<td>16.1</td>
<td>31.7</td>
<td>77.9</td>
</tr>
<tr>
<td>„Cho“</td>
<td>—</td>
<td>15.5</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>„Sutskever“</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATTN-ATTN</td>
<td>1</td>
<td>14.8</td>
<td>27.3</td>
<td>71.7</td>
</tr>
<tr>
<td>FINAL</td>
<td>—</td>
<td>10.8</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Selected models trained for translation from English to Czech. The embedding size is 1000 (except ATTN).
### Sample Results – translation quality en→cs

<table>
<thead>
<tr>
<th>Model Heads</th>
<th>Model</th>
<th>Heads</th>
<th>BLEU</th>
<th>Manual (&gt; other)</th>
<th>Manual (≥ other)</th>
</tr>
</thead>
<tbody>
<tr>
<td>„Bahdanau“</td>
<td>ATTN</td>
<td>—</td>
<td>22.2</td>
<td>50.9</td>
<td>93.8</td>
</tr>
<tr>
<td>compound attention</td>
<td>ATTN-ATTN</td>
<td>8</td>
<td>18.4</td>
<td>42.5</td>
<td>88.6</td>
</tr>
<tr>
<td></td>
<td>ATTN-ATTN</td>
<td>4</td>
<td>17.1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>inner attention + „Cho“</td>
<td>ATTN-CTX</td>
<td>4</td>
<td>16.1</td>
<td>31.7</td>
<td>77.9</td>
</tr>
<tr>
<td>„Cho“</td>
<td>FINAL-CTX</td>
<td>—</td>
<td>15.5</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>„Sutskever“</td>
<td>ATTN-ATTN</td>
<td>1</td>
<td>14.8</td>
<td>27.3</td>
<td>71.7</td>
</tr>
<tr>
<td></td>
<td>FINAL</td>
<td>—</td>
<td>10.8</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Heads</th>
<th>BLEU</th>
<th>Manual (&gt; other)</th>
<th>Manual (≥ other)</th>
</tr>
</thead>
<tbody>
<tr>
<td>„Bahdanau“</td>
<td>ATTN</td>
<td>—</td>
<td>22.2</td>
<td>50.9</td>
</tr>
<tr>
<td><strong>compound attention</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATTN-ATTN</td>
<td>8</td>
<td>18.4</td>
<td>42.5</td>
<td>88.6</td>
</tr>
<tr>
<td>ATTN-ATTN</td>
<td>4</td>
<td>17.1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>inner attention + „Cho“</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATTN-CTX</td>
<td>4</td>
<td>16.1</td>
<td>31.7</td>
<td>77.9</td>
</tr>
<tr>
<td>„Cho“</td>
<td>FINAL-CTX</td>
<td>—</td>
<td>15.5</td>
<td>—</td>
</tr>
<tr>
<td>** „Sutskever“**</td>
<td>ATTN-ATTN</td>
<td>1</td>
<td>14.8</td>
<td>27.3</td>
</tr>
<tr>
<td>FINAL</td>
<td>—</td>
<td>10.8</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

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

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

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

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

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

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

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

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

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

More heads → worse results.
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:

- $-0.57 \pm 0.31$ (en→cs)
- $-0.54 \pm 0.27$ (en→de)

Pairwise average (except BLEU):

- $0.78 \pm 0.32$ (en→cs)
- $0.62 \pm 0.23$ (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 data, no dose adjustment is necessary (see section 5.2).
Average attention weight by position

Relative position in encoder

Inner attention weight

Given the available clinical data, a k-ine-tic adjustment is necessary.
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 “meaning” representation.
- Heads divide sentence equidistantly, not logically.

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

\[
\mathcal{L} = \mathcal{L}_{MT} + \alpha \mathcal{L}_{MSE}
\]
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 $\alpha$ 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.


Bibliography


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.


Bibliography


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