Introduction to (Neural) Machine Translation

Ondřej Bojar

September 5, 2022
Outline

• Statistical Machine Translation
  • Why “classical” SMT fails
• Neural Machine Translation
  • Learning Representations
  • Processing Text
  • Transformer (Self-Attention)
• Some of Training Magic
• Caveats on Interpreting Results

Some slides by Jindřich Helcl, Jindřich Libovický, Tom Kocmi. Many slides based on slides by Rico Sennrich and others.
Further Reading

Slides from my subject here at ÚFAL:

https://ufal.mff.cuni.cz/courses/npl01087

Videolectures & Wiki (everything before neural):

http://mttalks.ufal.ms.mff.cuni.cz/

Books and others:


Chapter on NMT: https://arxiv.org/abs/1709.07809

• Ondřej Bojar: Čeština a strojový překlad. ÚFAL, 2012.
Approaches to Machine Translation

- The deeper analysis, the easier the **transfer** should be.
- A hypothetical interlingua captures pure meaning.
- Statistical systems learn “automatically” from data.
- Rule-based systems implemented by linguists-programmers.

Until NMT, it was best to combine the approaches.
The Statistical Approach

(Statistical \(=\) Information-theoretic.)

- Specify a probabilistic model.
  - How is the probability mass distributed among possible outputs given observed inputs.
- Specify the training criterion and procedure.
  - How to learn free parameters from training data.

Notice:

- Linguistics helpful when designing the models:
  - How to divide input into smaller units.
  - Which bits of observations are more informative.
Statistical MT

Given a source (foreign) language sentence $f_1^J = f_1 \ldots f_j \ldots f_J$, produce a target language (English) sentence $e_1^I = e_1 \ldots e_j \ldots e_I$. Among all possible target language sentences, choose the sentence with the highest probability:

$$\hat{e}_1^I = \arg\max_{I,e_1^I} p(e_1^I|f_1^J)$$

(1)

We stick to the $e_1^I, f_1^J$ notation regardless the source and target languages.
Brute-Force MT

Translate only sentences listed in a “translation memory” (TM):

Good morning.  =  Dobré ráno.
How are you?  =  Jak se máš?
How are you?  =  Jak se máte?

\[ p(e_1^I | f_1^J) = \begin{cases} 
1 & \text{if } e_1^I = f_1^J \text{ seen in the TM} \\
0 & \text{otherwise} 
\end{cases} \]  

(2)

Any problems with the definition?
Brute-Force MT

Translate only sentences listed in a “translation memory” (TM):

Good morning. = Dobré ráno.
How are you? = Jak se máš?
How are you? = Jak se máte?

\[
p(e^I_1 | f^J_1) = \begin{cases} 
1 & \text{if } e^I_1 = f^J_1 \text{ seen in the TM} \\
0 & \text{otherwise} 
\end{cases}
\]  

(2)

• Not a probability. There may be \( f^J_1 \), s.t. \( \sum_{e^I_1} p(e^I_1 | f^J_1) > 1 \).

\( \Rightarrow \) Have to normalize, use \( \frac{\text{count}(e^I_1, f^J_1)}{\text{count}(f^J_1)} \) instead of 1.

• Not “smooth”, no generalization:

Good morning. \( \Rightarrow \) Dobré ráno.
Good evening. \( \Rightarrow \) ∅
\[
\hat{e}_1^I = \arg\max_{I,e_1^I} p(e_1^I|f_1^J) = \arg\max_{I,e_1^I} p(f_1^J|e_1^I)p(e_1^I)
\] (3)

Bayes’ law divided the model into components:

- \(p(f_1^J|e_1^I)\) Translation model (“reversed”, \(e_1^I \rightarrow f_1^J\)) …is it a likely translation?
- \(p(e_1^I)\) Language model (LM) …is the output a likely sentence of the target language?

- The components can be trained on different sources.
  There are far more monolingual data \(\Rightarrow\) language model can be more reliable.
  ... smoothing enabled by a math trick.
Without Equations

Input Global Search
for sentence with highest probability

Output

Parallel Texts

Monolingual Texts

Translation Model

Language Model

Global Search
for sentence with highest probability
Translation Model of Phrase-Based MT

- The key element was phrase translation probability:
  \[
  p(\tilde{f}_k | \tilde{e}_k) = \frac{\text{count}(\tilde{f}, \tilde{e})}{\text{count}(\tilde{e})} \quad \text{... how often } \tilde{f}_k \text{ served as the translation of } \tilde{e}_k. 
  \]

- The whole source sentence was covered with phrase translations.

- The total translation probability was the product of individual phrase translation probabilities:
  \[
  h_{\text{Phr}}(f_1^J, e_1^I, s_1^K) = \log \prod_{k=1}^K p(\tilde{f}_k | \tilde{e}_k) 
  \]
  ... smoothing via decomposition into smaller units.
Work on Your Data

- CzEng (Czech-English) reached 180M million sentence pairs:
  - 0.6 cs / 0.7 en gigawords of genuine parallel text (61M sentpairs)
  - 2.0 cs / 2.3 en gigawords of synthetic text (127M sentpairs)

<table>
<thead>
<tr>
<th>Ver.</th>
<th>S. Pairs</th>
<th>Main Focus</th>
<th>Details in</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.9M</td>
<td>Sentence alignment, common format</td>
<td>Bojar and Žabokrtský (2006)</td>
</tr>
<tr>
<td>0.7</td>
<td>1.0M</td>
<td>Used in WMT06 and WMT07</td>
<td>Bojar et al. (2008)</td>
</tr>
<tr>
<td>0.9</td>
<td>8.0M</td>
<td>Automatic annotation up to t-layer</td>
<td>Bojar and Žabokrtský (2009)</td>
</tr>
<tr>
<td>–</td>
<td>–</td>
<td>Sentence-level filtering</td>
<td>Bojar et al. (2010)</td>
</tr>
<tr>
<td>1.0</td>
<td>15.0M</td>
<td>Improving monolingual annotation through parallel data</td>
<td>Bojar et al. (2012)</td>
</tr>
<tr>
<td>1.6</td>
<td>62.5M</td>
<td>Processing tools dockered</td>
<td>Bojar et al. (2016)</td>
</tr>
<tr>
<td>1.7</td>
<td>57.1M</td>
<td>Block-level filtering</td>
<td>–</td>
</tr>
<tr>
<td>2.0</td>
<td>188.0M</td>
<td>Filtering + Synthetic data</td>
<td>–</td>
</tr>
</tbody>
</table>
In my dream, there was a sycamore growing out of the ruins of the sacristy, and I was told that, if I dug at the roots of the sycamore, I would find a hidden treasure. But I’m not so stupid as to cross an entire desert just because of a recurrent dream. "And they disappeared. The boy stood up shakily, and looked once more at the Pyramids. "It is I who dared to do so," said the boy. This man looked exactly the same, except that now the roles were reversed. "It is I who dared to do so," he
In my dream, there was a sycamore growing out of the ruins of the sacristry, and I was told that, if I dug at the roots of the sycamore, I would find a hidden treasure.

But I’m not so stupid as to cross an entire desert just because of a recurrent dream.

And they disappeared.

The boy stood up shakily, and looked once more at the Pyramids.

"It is I who dared to do so," said the boy.

This man looked exactly the same, except that now the rope was reversed.

"It is so who dared to do so," he repeated, and he lowered his head to receive a blow from the sword.

"Life was good to me," the man said.

“When you appeared in my dream, I felt that all my efforts had been rewarded, because my son’s poems will be read by men for generations to come.

But any father would be proud of the fame achieved by one whom he had cared for as a child, and educated as he grew up.

We’re two very different things.

"That’s not true," the boy said.

I have inside me the winds, the deserts, the oceans, the stars, and everything created in the universe.

We were all made by the same hand, and we have the same soul.

You’ll learn to love the desert, and you’ll get to know every one of the fifty thousand palms.

You’ll watch them as they grow, demonstrating how the world is always changing.

And you’ll get better and better at understanding ommens, because the desert is the best teacher there is.

"Sometimes during the second year, you’ll remember about the treasure.

The ommens will begin insistently to speak of it, and you’ll try to ignore them.

But you know that I’m not going to go to Mecca. Just as you know that you’re not going to buy your sheep.

“Who told you that?” asked the boy, startled.

“Makathub” said the old crystal merchant.

And he gave the boy his blessing.

The boy went to his room and packed his belongings.

They filled three sacks.

As he was leaving, he saw, in the corner of the room, his old shepherd’s soucha.

"I want to see the greatness of Allah,” the chief said, with respect.

“I want to see how a man turns himself into the wind.

But he made a mental note of the names of the two men who had expressed their fear.

Gale and Church (1993) illustrated in MT Talk #7 (https://youtu.be/_4lnyoC3mtQ)
From Sent. Pairs, Learn Word Translations

IBM Model 1 illustrated in Bojar (2012).
See also MT Talk #8 (https://youtu.be/mqyMDLuj5JPw)
Nemám žádného psa.  Viděl kočku.
I have no dog.  He saw a cat.
Nemám žádného psa.
I have no dog.

Viděl kočku.
He saw a cat.
Nemám žádného psa.
I have no dog.
Viděl kočku.
He saw a cat.
Nemám žádného psa.
I have no dog.
Viděl kočku.
He saw a cat.

New input: Nemám kočku.
New input: Nemám kočku. ... I don't have cat.

Nemám žádného psa.
I have no dog.

Viděl kočku.
He saw a cat.
Nemám žádného psa.
I have no dog.
Viděl kočku.
a cat. He saw
New input: kočku.
Nemám...
I don't have cat.
New input: Nemám
I have kočku.
6: So That $n$-Grams Probable (LM)

Nemám žádného psa. I have no dog. Viděl kočku. He saw a cat.

New input: Nemám kočku. I don't have cat.
Nemám žádného psa.
I have no dog.
Viděl kočku.
a cat. He saw a cat.

New input: Nemám kočku.
I don't have a cat.
What Went Wrong?

\[
\hat{e}_{1}^{I} = \arg\max_{I, e_{1}^{I}} p(f_{1}^{J}|e_{1}^{I}) p(e_{1}^{I}) = \arg\max_{I, e_{1}^{I}} \prod_{(f, \hat{e}) \in \text{phrase pairs of } f_{1}^{J}, e_{1}^{I}} p(f|\hat{e}) p(e_{1}^{I})
\]

(4)

Too strong independence assumptions:

- **Language model as a separate unit.**
  - \( p(e_{1}^{I}) \) models the target sentence independently of \( f_{1}^{J} \).
- **Phrases translated independent of one another.**
  - In fact, phrases do depend on each other.
    Here “nemám” and “žádného” jointly express one negation.
  - Word alignments ignored that dependence.
    But adding it would increase data sparseness.
  - LM was the only means for glueing phrases, but it prefers positive sents.
Redefining \( p(e_1^I|f_1^J) \)

What if we modelled \( p(e_1^I|f_1^J) \) directly, word by word:

\[
p(e_1^I|f_1^J) = p(e_1, e_2, \ldots, e_I|f_1^J) \\
= p(e_1|f_1^J) \cdot p(e_2|e_1, f_1^J) \cdot p(e_3|e_2, e_1, f_1^J) \cdots \\
= \prod_{i=1}^{I} p(e_i|e_1, \ldots, e_{i-1}, f_1^J)
\]

(5)

…this is “just a cleverer language model:” \( p(e_1^I) = \prod_{i=1}^{I} p(e_i|e_1, \ldots, e_{i-1}) \)

Main Benefit: All dependencies available.

But what technical device can learn this?
• A neural network with a single hidden layer (possibly huge) can approximate any continuous function to any precision.
• (Nothing claimed about learnability in practice.)

$-0.43x_1 - 0.89x_2 + 2.0 > 0$
and $-0.67x_1 + 0.89x_2 + 2.1 > 0$
and $1.4x_1 - 0.067x_2 + 2.3 > 0$
A DL “Program” Is Just a Computation…

In fact:  
\[ 1 \tanh(-0.43x_1 - 0.89x_2 + 2.0) + 1 \tanh(-0.67x_1 + 0.89x_2 + 2.1) + 1 \tanh(1.4x_1 - 0.067x_2 + 2.3) - \pi/2 > 0 \]
... with Parameters Guessed Automatically

In fact: \( 1 \tanh(-0.43x_1 - 0.89x_2 + 2.0) + 1 \tanh(-0.67x_1 + 0.89x_2 + 2.1) + 1 \tanh(1.4x_1 - 0.067x_2 + 2.3) - \pi/2 > 0 \)
Perfect Features

\[ 1x_1^2 + 1x_2^2 - 1 < 0 \]
Bad Features & Low Depth
Too Complex NN Fails to Learn

This is the output from one neuron. Hover to see it larger.

The outputs are mixed with varying weights, shown by the thickness of the lines.
One Fully Connected Layer

- One fully-connected layer converts an input (column) vector $x$ to an output (column) vector $h$:

$$h = f(Wx + b),$$

- $W$ is a weight matrix of input columns and output rows,
- $b$ a bias vector of length of output,
- $f(\cdot)$ is a non-linearity applied usually elementwise.
One Layer $\tanh(Wx + b)$, 2D $\rightarrow$ 2D

Skew: $W$

Transpose: $b$

Non-lin.: $\tanh$

Feed-Forward Neural Network

\[ x \]
\[ h_1 = f(W_1 x + b_1) \]
\[ h_2 = f(W_2 h_1 + b_2) \]
\[ \vdots \]
\[ h_n = f(W_n h_{n-1} + b_n) \]
\[ o = g(W_o h_n + b_o) \]
\[ E = e(o, t) \]

\[ \frac{\partial E}{\partial W_o} = \frac{\partial E}{\partial o} \cdot \frac{\partial o}{\partial W_o} \]

**blue**: Training item (input \( x \), output \( t \)), **red**: Trainable parameters \((W_1, b_1, \ldots)\).
Four Layers, Disentangling Spirals

Deep NNs for Image Classification

It's deep if it has more than one stage of non-linear feature transformation
Processing Text with NNs

- Map each word to a vector of 0s and 1s ("1-hot repr."): cat \(\rightarrow (0, 0, \ldots, 0, 1, 0, \ldots, 0)\)

- Sentence is then a matrix:

\[
\begin{array}{cccccc}
\text{the} & \text{cat} & \text{is} & \text{on} & \text{the} & \text{mat} \\
\hline
a & 0 & 0 & 0 & 0 & 0 \\
about & 0 & 0 & 0 & 0 & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
cat & 0 & 1 & 0 & 0 & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
is & 0 & 0 & 1 & 0 & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
the & 1 & 0 & 0 & 0 & 1 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\hline
zebra & 0 & 0 & 0 & 0 & 0
\end{array}
\]

Main drawback: No relations, all words equally close/far.
• Map each word to a vector of 0s and 1s ("1-hot repr."): 
\[ \text{cat} \mapsto (0, 0, \ldots, 0, 1, 0, \ldots, 0) \]

• Sentence is then a matrix:

<table>
<thead>
<tr>
<th></th>
<th>the</th>
<th>cat</th>
<th>is</th>
<th>on</th>
<th>the</th>
<th>mat</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>a</code></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><code>about</code></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><code>cat</code></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><code>zebra</code></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Vocabulary size:
- 1.3M English
- 2.2M Czech

Main drawback: No relations, all words equally close/far.
### Processing Text with NNs

- **Map each word to a vector of 0s and 1s (“1-hot repr.”):**
  
  \[
  \text{cat} \mapsto (0, 0, \ldots, 0, 1, 0, \ldots, 0)
  \]

- **Sentence is then a matrix:**

  \[
  \begin{array}{ccccccc}
  & \text{the} & \text{cat} & \text{is} & \text{on} & \text{the} & \text{mat} \\
  \uparrow & a & 0 & 0 & 0 & 0 & 0 \\
  & \text{about} & 0 & 0 & 0 & 0 & 0 \\
  & \text{...} & \text{...} & \text{...} & \text{...} & \text{...} & \text{...} \\
  & \text{cat} & 0 & 1 & 0 & 0 & 0 \\
  \text{Vocabulary size:} & \text{...} & \text{...} & \text{...} & \text{...} & \text{...} & \text{...} \\
  \text{1.3M English} & \text{is} & 0 & 0 & 1 & 0 & 0 \\
  \text{2.2M Czech} & \text{...} & \text{...} & \text{...} & \text{...} & \text{...} & \text{...} \\
  & \text{the} & 1 & 0 & 0 & 0 & 1 \\
  & \text{...} & \text{...} & \text{...} & \text{...} & \text{...} & \text{...} \\
  \downarrow & \text{zebra} & 0 & 0 & 0 & 0 & 0 \\
  \end{array}
  \]

  **Main drawback:** No relations, all words equally close/far. *(Smoothing!)*
• Idea: Map each word to a dense vector.
• Result: 300–2000 dimensions instead of 1–2M.
  • The dimensions have no clear interpretation.
• The “embedding” is the mapping.
  • Technically, the first layer of NNs for NLP is the matrix that maps 1-hot input to the first layer.

• Embeddings are trained for each particular task.
  • Sentence classification (sentiment analysis, etc.)
  • Neural language modelling.
  • The famous word2vec (Mikolov et al., 2013):
    • CBOW: Predict the word from its four neighbours.
    • Skip-gram: Predict likely neighbours given the word.
  • End-to-end neural MT.
2: Reducing Voc. Size: Sub-Words

- SMT struggled with productive morphology (>1M wordforms).
  - nejneobhodpodávávatelnějšími, Donaudampfschiffahrtsgesellschaftskapitän
- NMT can handle vocabulary of only only 30–80k entries.

⇒ Resort to sub-word units.

<table>
<thead>
<tr>
<th></th>
<th>český politik svezl migranty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syllables</td>
<td>čes ký po li tik sve zl mig ran ty</td>
</tr>
<tr>
<td>Morphemes</td>
<td>česk ý politik vez l migrant y</td>
</tr>
<tr>
<td>Char Pairs</td>
<td>če sk ý po li ti k sv ez l mi gr an ty</td>
</tr>
<tr>
<td>Chars</td>
<td>č e s k ý p o l i t i k s v e z l m i g r a n t y</td>
</tr>
<tr>
<td>BPE 30k</td>
<td>český politik vez vez mi granty</td>
</tr>
</tbody>
</table>

BPE (Byte-Pair Encoding, (Sennrich et al., 2016)) or Google’s wordpieces (Wu et al., 2016) and Tensor2Tensor’s SubwordTextEncoder use \( n \) most common substrings (incl. frequent words).

Variable-length input can be handled by recurrent NNs:

- Processing one input symbol at a time.
  - Initial state $h_0 = (0)$ (or some sentence representation).
  - The same (trained) transformation $A$ used every time.

$$h_t = A(h_{t-1}, x_t)$$

- Unroll in time (up to a fixed length limit).
Vanilla RNN

\[ h_t = \tanh(W[h_{t-1}; x_t] + b) \]  

- Vanishing gradient problem.
- Non-linear transformation always applied.
  \[ [h_{t-1}; x_t] \] is concatenation of \( h_{t-1} \) and \( x_t \)
  \( \Rightarrow \) Type theory: \( h_t \) and \( h_{t-1} \) live in different vector spaces.
LSTM and GRU Cells for RNN

- LSTM, Long Short-Term Memory Cells (Hochreiter and Schmidhuber, 1997).
- GRU, Gated Recurrent Unit Cells (Chung et al., 2014):

\[
\begin{align*}
    z_t &= \sigma (W_z [h_{t-1}; x_t] + b_z) \quad (9) \\
    r_t &= \sigma (W_r [h_{t-1}; x_t] + b_r) \quad (10) \\
    \tilde{h}_t &= \tanh (W [r_t \odot h_{t-1}; x_t]) \quad (11) \\
    h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (12)
\end{align*}
\]

- Gates control:
  - what to use from input \( x_t \) (GRU: everything),
  - what to use from hidden state \( h_{t-1} \) (reset gate \( r_t \)),
  - what to put into output (update gate \( z_t \))

- Linear “information highway” preserved.
  \( \Rightarrow \) All states \( h_t \) belong to the same vector space.
Cho et al. (2014) proposed:

- encoder-decoder architecture and
- GRU unit (name given later by Chung et al. (2014))
- to score variable-length phrase pairs in PBMT.
Embeddings of Phrases
RNN Language Model

- Train RNN as a **classifier for next words** (unlimited history):

  \[
  \langle s \rangle \downarrow w_1 \downarrow w_2 \downarrow w_3 \downarrow w_4 \downarrow \vdots
  \]

  \[
  p(w_1) \quad p(w_2) \quad p(w_3) \quad p(w_4) \quad p(w_5)
  \]

- Can be used:
  - To estimate sentence probability / perplexity.
  - To sample from the distribution:
Two Views on RNN LM

• RNN is a for loop / functional map over sequential data
• all outputs are conditional distributions
  → probabilistic distribution over sequences of words

\[ P (w_1, \ldots, w_n) = \prod_{i=1}^{n} P (w_i | w_{i-1}, \ldots, w_1) \]
RNN for Translation: Encoder-Decoder

(source language input + target language LM)
RNN for Translation: Encoder-Decoder

(source language input + target language LM)
Encoder-Decoder Model – Formal Notation

**Data**

input tokens (source language) \( \mathbf{x} = (x_1, \ldots, x_{T_x}) \)

output tokens (target language) \( \mathbf{y} = (y_1, \ldots, y_{T_y}) \)
Data
input tokens (source language) \( x = (x_1, \ldots, x_{T_x}) \)
output tokens (target language) \( y = (y_1, \ldots, y_{T_y}) \)

Encoder
initial state \( h_0 \equiv 0 \)
j-th state \( h_j = \text{RNN}_{\text{enc}}(h_{j-1}, x_j) = \tanh(U_e h_{j-1} + W_e E_e x_j + b_e) \)
final state \( h_{T_x} \)
Encoder-Decoder Model – Formal Notation

Data
input tokens (source language) \( x = (x_1, \ldots, x_{T_x}) \)
output tokens (target language) \( y = (y_1, \ldots, y_{T_y}) \)

Encoder
initial state \( h_0 \equiv \mathbf{0} \)
\( j \)-th state \( h_j = \text{RNN}_{\text{enc}}(h_{j-1}, x_j) = \tanh(U_e h_{j-1} + W_e E_e x_j + b_e) \)
final state \( h_{T_x} \)

Decoder
initial state \( s_0 = h_{T_x} \)
\( i \)-th decoder state \( s_i = \text{RNN}_{\text{dec}}(s_{i-1}, \hat{y}_{i-1}) = \tanh(U_d s_{i-1} + W_d E_d \hat{y}_{i-1} + b_d) \)
\( i \)-th word score \( t_i = \tanh(U_o s_i + W_o E_d \hat{y}_{i-1} + b_o) \) (“output projection”)
output \( \hat{y}_i = \arg \max V_o t_i \)
Implementation: Training vs. Runtime

**Training:** $y_j$ (ground truth)  

**Runtime:** $\hat{y}_j$ (decoded)
Summary So Far

- Statistical MT chooses the most probable sentence:
  \[ \hat{e}_I^1 = \arg\max_{I, e_I^1} p(e_I^1 | f_J^1) \]
- Independence assumptions (LM vs. TM; phrase independence) were harmful.
- Neural MT predicts word by word; “just a clever LM".
  \[ p(e_I^1 | f_J^1) = \prod_{i=1}^{I} p(e_i | e_1, \ldots, e_{i-1}, f_J^1) \]
  - Sub-word units, word embeddings, RNN for variable-length, encoder-decoder.
- We moved from searching for best minimum translation units to representing words and phrases in continuous space and relating them to each other.
Attention is All You Need (Vaswani et al., 2017)
Transformer Detailed Walkthroughs

Transformer Illustrated:
- http://jalammar.github.io/illustrated-transformer/
  Amazingly simple description! (I am reusing the pictures.)

Transformer paper annotated with PyTorch code:
- http://nlp.seas.harvard.edu/2018/04/03/attention.html
- PyTorch by examples:
  https://github.com/jcjohnson/pytorch-examples

Summary at Medium:
- https://medium.com/@adityathiruvengadam/
  transformer-architecture-attention-is-all-you-need-aeccd9f50d09
Self-Attention Motivation

- Sequences of arbitrary length $n$ need to be processed.
- Information gets lost over too many processing steps.
- RNNs make the (time-unrolled) network as deep as $n$.
- CNNs allow to trade kernel size $k$ and depth for a target "receptive field":

![Diagram showing self-attention mechanism]

- SANs (Self-Attentive Networks) can access any position in constant time.
Self-Attention

- Goal: **Aggregate** arbitrary-length input to fixed-size vector. Allow **data-driven, trainable** aggregation.
Self-Attention

- Goal: **Aggregate** arbitrary-length input to fixed-size vector. Allow **data-driven, trainable** aggregation.

Given the sequence of inputs \( x_1, \ldots, x_n \):

- Create three “views” of them: queries, keys, values.
- Using trained matrices \( W^Q, W^K, W^V \).
Match All Queries with All Keys

Input

Embedding

Queries

Keys

Values

Score

x_1

q_1

k_1

v_1

q_1 \cdot k_1 = 112

x_2

q_2

k_2

v_2

q_1 \cdot k_2 = 96
Normalize Scores

- **Input**
  - **Embedding**
  - **Values**
  - **Score**
    - Divide by $8 (\sqrt{d_k})$
  - **Softmax**

**Thinking**
- $x_1$
- $v_1$
- $q_1 \cdot k_1 = 112$
- $14$
- $0.88$

**Machines**
- $x_2$
- $v_2$
- $q_1 \cdot k_2 = 96$
- $12$
- $0.12$
Aggregate Values Accordingly

Input

Embedding

Values

Softmax

Softmax X Value

Sum

Thinking

Machines

x₁

v₁

v₂

0.88

x₂

v₁

v₂

0.12

z₁

z₂
Transformer = 6 Layers Enc + 6 Dec

INPUT: Je suis étudiant

OUTPUT: I am a student
Composition of One Layer
Self-Attention in Transformer

Three uses of multi-head attention in Transformer

- **Encoder-Decoder Attention:**
  - Q: previous decoder layers; K = V: outputs of encoder
  - Decoder positions attend to all positions of the input.

- **Encoder Self-Attention:**
  - Q = K = V: outputs of the previous layer of the encoder
  - Encoder positions attend to all positions of previous layer.

- **Decoder Self-Attention:**
  - Q = K = V: outputs of the previous decoder layer.
  - Masking used to prevent depending on future outputs.
  - Decoder attends to all its previous outputs.
Multi-Head Attention

Calculating attention separately in eight different attention heads

ATTENTION HEAD #0
Z₀

ATTENTION HEAD #1
Z₁

ATTENTION HEAD #7
Z₇
Self-Attention at Enc Layer #5: 1 Head
Self-Attention at Enc Layer #5: 2 Heads
Self-Attention at Enc Layer #5: 8 Heads
Some of Training Magic
Kocmi and Bojar (2017) explore curriculum learning:
- Start with simpler sentences first, add complex ones later.

For further important insights about sentence length overfitting, see Variš and Bojar (2021).
Curriculum vs. Catastrophic Forgetting

- Kocmi and Bojar (2017) explore curriculum learning:
  - Start with simpler sentences first, add complex ones later.
  - When “simpler” means “shorter”:

For further important insights about sentence length overfitting, see Variš and Bojar (2021).
• Kocmi and Bojar (2017) explore curriculum learning:
  • Start with simpler sentences first, add complex ones later.
• When “simpler” means “shorter”:
  • Clear jumps in score as bins of longer sentences are allowed.

For further important insights about sentence length overfitting, see Variš and Bojar (2021).
Kocmi and Bojar (2017) explore curriculum learning:
- Start with simpler sentences first, add complex ones later.
- When "simpler" means "shorter":
  - Clear jumps in score as bins of longer sentences are allowed.
  - Reversed curriculum unlearns to produce long sentences.

For further important insights about sentence length overfitting, see Variš and Bojar (2021).
See Popel et al. (2020) for more details.
• WMT 2018 English-to-Czech news translation results:

<table>
<thead>
<tr>
<th></th>
<th>Ave. %</th>
<th>Ave. z</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>84.4</td>
<td>0.667</td>
<td><strong>CUNI-Transformer</strong></td>
</tr>
<tr>
<td>2</td>
<td>79.8</td>
<td>0.521</td>
<td><strong>UEDIN</strong></td>
</tr>
<tr>
<td></td>
<td>78.6</td>
<td>0.483</td>
<td><strong>Professional Translation</strong></td>
</tr>
<tr>
<td>3</td>
<td>68.1</td>
<td>0.128</td>
<td><strong>ONLINE-B</strong></td>
</tr>
<tr>
<td>4</td>
<td>59.4</td>
<td>−0.178</td>
<td><strong>ONLINE-A</strong></td>
</tr>
<tr>
<td>5</td>
<td>54.1</td>
<td>−0.354</td>
<td><strong>ONLINE-G</strong></td>
</tr>
</tbody>
</table>

Caveats:
• Humans translated whole documents, MT individual segments.
• Evaluation was done for individual segments.


• WMT 2018 English-to-Czech news translation results:

<table>
<thead>
<tr>
<th>Ave. %</th>
<th>Ave. z</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>84.4</td>
<td>0.667</td>
<td>CUNI-Transformer</td>
</tr>
<tr>
<td>79.8</td>
<td>0.521</td>
<td>UEDIN</td>
</tr>
<tr>
<td>78.6</td>
<td>0.483</td>
<td>Professional Translation</td>
</tr>
<tr>
<td>68.1</td>
<td>0.128</td>
<td>ONLINE-B</td>
</tr>
<tr>
<td>59.4</td>
<td>−0.178</td>
<td>ONLINE-A</td>
</tr>
<tr>
<td>54.1</td>
<td>−0.354</td>
<td>ONLINE-G</td>
</tr>
</tbody>
</table>

Caveats:
• Humans translated whole documents, MT individual segments.
  • Evaluation was done for individual segments.
ARTICLE

Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals

Martin Popel, Marketa Tomkova, Jakub Tomek, Lukasz Kaiser, Jakob Uszkoreit, Ondrej Bojar & Zdenek Zabokrtsky
Manual evaluation by domain experts, scoring in categories:

1. Language Resources – Spelling and Morphology
2. Vocabulary – Adequacy of Terms Used
3. Vocabulary – Clarity of the Text in Terms of Used Words
4. Syntax and Word Order
5. Coherence and Overall Understanding of the Text

plotted as average rank for better comparibility
Supplement No. 1 to the agreement on the sublease the apartment, of 13th May 2016
On the day, month and year written below Marta Burešová, pers. no. 695604/3017
Address: Radimova 8, Prague 6, 169 00 as the tenant on the one hand (Hereinafter referred to as ”the tenant”) and Karolína Černá, pers. no. 136205/891 Address: Alfrédova 13, Praha 4, 142 00 As a lessee on the other (Hereinafter referred to as ”the lessee”) collectively also referred to as ”the Contracting parties” have agreed on this Supplement No. 1 to the Agreement on the sublease the apartment, of 13th May 2016 (hereinafter referred to as the ”Supplement No. 1”)
I. Introductory Provisions
On 13th May 2016, the tenant and the lessee closed the Agreement on the sublease of the apartment, under which the tenant let the lessee use the apartment No. 4 (area 49 m²) of size 1+1/L in the ground floor of the house in Prague 4, Alfrédova 13, ...
Dodatek č. 1 ke smlouvě o podnájmu bytu ze dne 13. května 2016
V den, měsíc a rok níže napsané Marta Burešová, pers. no. 695604/3017 Adresa: Radimova 8, Praha 6, 169 00 jako nájemce na jedné straně (dále jen “nájemce”) a Karolína Černá, pers. no. 136205/891 Adresa: Alfrédova 13, Praha 4, 142 00 jako nájemce na straně druhé (dále jen “nájemce”) společně označované také jako “smluvní strany” se dohodly na tomto dodatku č. 1 ke smlouvě o podnájmu, dále jen “nájemní smlouva”, dále jen “13. května 2016”).
I. Úvodní ustanovení
Dne 13. května 2016 nájemce a nájemce uzavřeli smlouvu o dalším pronájmu bytu, podle níž nájemce pronajímá nájemci byt č. 4 (plocha 49 m²) o velikosti 1+1/l v přízemí domu v Praze 4, Alfrédova 13, ...
Dodatek č. 1 ke smlouvě o podnájmu bytu ze dne 13. května 2016
V den, měsíc a rok níže napsané Marta Burešová, pers. no. 695604/3017 Adresa: Radimova 8, Praha 6, 169 00 jako nájemce na jedné straně (dále jen “nájemce”) a Karolína Černá, pers. no. 136205/891 Adresa: Alfrédova 13, Praha 4, 142 00 jako nájemce na straně druhé (dále jen “nájemce”) společně označované také jako “smluvní strany” se dohodly na tomto dodatku č. 1 ke smlouvě o podnájmu, dále jen “nájemní smlouva”, dále jen “13. května 2016”).
I. Úvodní ustanovení
Dne 13. května 2016 nájemce a nájemce uzavřeli smlouvu o dalším pronájmu bytu, podle níž nájemce pronajímá nájemci byt č. 4 (plocha 49 m²) o velikosti 1+1/l v přízemí domu v Praze 4, Alfrédova 13, …
Caveats on Interpreting Results
“Modeling Source Syntax Helps NMT” (1/2)

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>36.66</td>
<td>–</td>
</tr>
<tr>
<td>Parse from layer 0</td>
<td>36.60</td>
<td>82.85</td>
</tr>
<tr>
<td>Parse from layer 1</td>
<td><strong>38.01</strong></td>
<td>90.78</td>
</tr>
<tr>
<td>Parse from layer 2</td>
<td>37.87</td>
<td>91.18</td>
</tr>
<tr>
<td>Parse from layer 3</td>
<td>37.67</td>
<td>91.43</td>
</tr>
<tr>
<td>Parse from layer 4</td>
<td>37.60</td>
<td><strong>91.56</strong></td>
</tr>
<tr>
<td>Parse from layer 5</td>
<td>37.67</td>
<td>91.46</td>
</tr>
</tbody>
</table>

• Forcing one Trafo head to provide dependency tree helps BLEU.

"Modeling Source Syntax Helps NMT" (1/2)

- Forcing one Trafo head to provide dependency tree helps BLEU.
- Forcing one Trafo head to provide linear tree helps more.

Alternating output words and CCG tags helps. (Nadejde et al. 2017)

Tgt: NP Obama ((S[dcl]\NP)/PP)/NP receives NP Net+ an+ yahu PP/NP in NP/N the N capital (NP/NP/NP)/NP

We tried the same with: (RNN or Transformer; interleaved or multi-decoder)
- correct CCG tags, • random tags, ⋆ a single dummy tag.

Except ⋆ single dummy tag, both improve over the • baseline.

Transfer Learning (TL)

Parent corpus

Child corpus

See Kocmi and Bojar (2018) for more details.
Transfer Learning (TL)

Parent corpus

Child corpus

Balanced vocabulary

See Kocmi and Bojar (2018) for more details.
Transfer Learning (TL)

Parent corpus

CS
EN

Child corpus

ET
EN

Balanced vocabulary

See Kocmi and Bojar (2018) for more details.
Child model: Slovak

<table>
<thead>
<tr>
<th>Parent model</th>
<th>Corpus size difference</th>
<th>Direction</th>
<th>Baseline (BLEU)</th>
<th>Transfer (BLEU)</th>
<th>Δ (BLEU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>9x from English</td>
<td>16.13</td>
<td>17.75</td>
<td>1.62 *</td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td>9x to English</td>
<td>19.19</td>
<td>22.42</td>
<td>3.23 *</td>
<td></td>
</tr>
</tbody>
</table>

* statistically significant

See Kocmi and Bojar (2018) for more details.
**Child model: Slovak**

<table>
<thead>
<tr>
<th>Parent model</th>
<th>Corpus size difference</th>
<th>Direction</th>
<th>Baseline (BLEU)</th>
<th>Transfer (BLEU)</th>
<th>Δ (BLEU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>9x</td>
<td>from English</td>
<td>16.13</td>
<td>17.75</td>
<td>1.62 *</td>
</tr>
<tr>
<td>Czech</td>
<td>9x</td>
<td>to English</td>
<td>19.19</td>
<td>22.42</td>
<td>3.23 *</td>
</tr>
</tbody>
</table>

* statistically significant

See Kocmi and Bojar (2018) for more details.
## “TL Exploits Language Similarity”

### Child model: Slovak

<table>
<thead>
<tr>
<th>Parent model</th>
<th>Corpus size difference</th>
<th>Direction</th>
<th>Baseline (BLEU)</th>
<th>Transfer (BLEU)</th>
<th>Δ (BLEU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>9x</td>
<td>from English</td>
<td>16.13</td>
<td>17.75</td>
<td>1.62 *</td>
</tr>
<tr>
<td>Czech</td>
<td>9x</td>
<td>to English</td>
<td>19.19</td>
<td>22.42</td>
<td>3.23 *</td>
</tr>
</tbody>
</table>

### Child model: Estonian

<table>
<thead>
<tr>
<th>Parent model</th>
<th>Corpus size difference</th>
<th>Direction</th>
<th>Baseline (BLEU)</th>
<th>Transfer (BLEU)</th>
<th>Δ (BLEU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finnish</td>
<td>3.5x</td>
<td>from English</td>
<td>17.03</td>
<td>19.74</td>
<td>2.71 *</td>
</tr>
<tr>
<td>Russian</td>
<td>16x</td>
<td>from English</td>
<td>17.03</td>
<td>20.09</td>
<td>3.06 *</td>
</tr>
<tr>
<td>Czech</td>
<td>50x</td>
<td>from English</td>
<td>17.03</td>
<td>20.41</td>
<td>3.38 *</td>
</tr>
<tr>
<td>Finnish</td>
<td>3.5x</td>
<td>to English</td>
<td>21.74</td>
<td>24.18</td>
<td>2.44 *</td>
</tr>
<tr>
<td>Russian</td>
<td>16x</td>
<td>to English</td>
<td>21.74</td>
<td>23.54</td>
<td>1.80 *</td>
</tr>
</tbody>
</table>

* statistically significant

See Kocmi and Bojar (2018) for more details.
### TL Exploits Language Similarity

#### Child model: Slovak

<table>
<thead>
<tr>
<th>Parent model</th>
<th>Corpus size difference</th>
<th>Direction</th>
<th>Baseline (BLEU)</th>
<th>Transfer (BLEU)</th>
<th>Δ (BLEU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>9x</td>
<td>from English</td>
<td>16.13</td>
<td>17.75</td>
<td>1.62 *</td>
</tr>
<tr>
<td>Czech</td>
<td>9x</td>
<td>to English</td>
<td>19.19</td>
<td>22.42</td>
<td>3.23 *</td>
</tr>
</tbody>
</table>

#### Child model: Estonian

<table>
<thead>
<tr>
<th>Parent model</th>
<th>Corpus size difference</th>
<th>Direction</th>
<th>Baseline (BLEU)</th>
<th>Transfer (BLEU)</th>
<th>Δ (BLEU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finnish</td>
<td>3.5x</td>
<td>from English</td>
<td>17.03</td>
<td>19.74</td>
<td>2.71 *</td>
</tr>
<tr>
<td>Russian</td>
<td>16x</td>
<td>from English</td>
<td>17.03</td>
<td>20.09</td>
<td>3.06 *</td>
</tr>
<tr>
<td>Czech</td>
<td>50x</td>
<td>from English</td>
<td>17.03</td>
<td>20.41</td>
<td>3.38 *</td>
</tr>
<tr>
<td>Finnish</td>
<td>3.5x</td>
<td>to English</td>
<td>21.74</td>
<td>24.18</td>
<td>2.44 *</td>
</tr>
<tr>
<td>Russian</td>
<td>16x</td>
<td>to English</td>
<td>21.74</td>
<td>23.54</td>
<td>1.80 *</td>
</tr>
</tbody>
</table>

* statistically significant

See Kocmi and Bojar (2018) for more details.
### TL Exploits Language Similarity

**Child model: Slovak**

<table>
<thead>
<tr>
<th>Parent model</th>
<th>Corpus size difference</th>
<th>Direction</th>
<th>Baseline (BLEU)</th>
<th>Transfer (BLEU)</th>
<th>Δ (BLEU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Related</td>
<td>9x</td>
<td>from English</td>
<td>16.13</td>
<td>17.75</td>
<td>1.62 *</td>
</tr>
<tr>
<td>Related</td>
<td>9x</td>
<td>to English</td>
<td>19.19</td>
<td>22.42</td>
<td>3.23 *</td>
</tr>
</tbody>
</table>

**Child model: Estonian**

<table>
<thead>
<tr>
<th>Parent model</th>
<th>Corpus size difference</th>
<th>Direction</th>
<th>Baseline (BLEU)</th>
<th>Transfer (BLEU)</th>
<th>Δ (BLEU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Related</td>
<td>3.5x</td>
<td>from English</td>
<td>17.03</td>
<td>19.74</td>
<td>2.71 *</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>16x</td>
<td>from English</td>
<td>17.03</td>
<td>20.09</td>
<td>3.06 *</td>
</tr>
<tr>
<td>Biggest</td>
<td>50x</td>
<td>from English</td>
<td>17.03</td>
<td>20.41</td>
<td>3.38 *</td>
</tr>
<tr>
<td>Related</td>
<td>3.5x</td>
<td>to English</td>
<td>21.74</td>
<td>24.18</td>
<td>2.44 *</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>16x</td>
<td>to English</td>
<td>21.74</td>
<td>23.54</td>
<td>1.80 *</td>
</tr>
</tbody>
</table>

* statistically significant

See Kocmi and Bojar (2018) for more details.
Summary

• Given all the data we had for SMT, given big GPUs, given all the training tricks, and given a few weeks of training time, Transformer can reach (average) quality comparable to (sloppy) humans.

• Intuition why something works is often wrong.
  • Use trivial baselines to exclude misinterpretations; read more here:

References

Ondřej Bojar, Miroslav Janíček, Zdeněk Žabokrtský, Pavel Češka, and Peter Beňa. 2008. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In Proceedings of the Sixth International Language Resources and Evaluation (LREC’08), Marrakech, Morocco, May. ELRA.


