Extracting Syntactic Trees from NMT Encoder Self-Attentions

David Mareček and Rudolf Rosa
Charles University, Institute of Formal and Applied Linguistics
{marecek,rosa}@ufal.mff.cuni.cz

CURRENT BEST METHODS FOR MACHINE TRANSLATION DO NOT USE ANY LINGUISTIC ANNOTATIONS

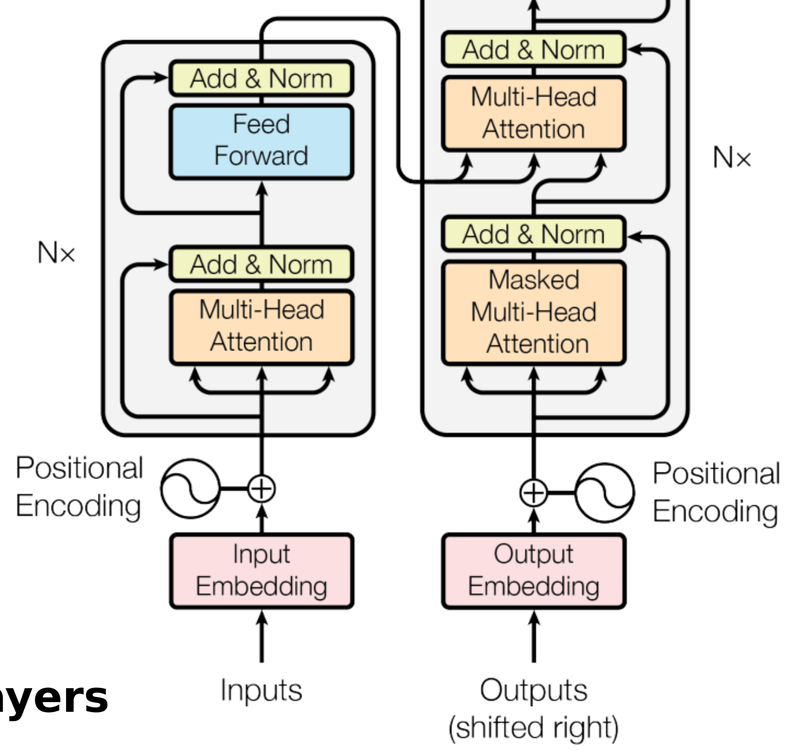
IS THERE ANY LANTENT SYNTAX LEARNED BY NEURAL MACHINE TRANSLATION ???

Goals

Use state-of-the art NMT system Transformer
Extract and analyze encoder self-attentions
Create parse trees using self-attentions matrices
Compare it to manually annotated treebanks

Transformer

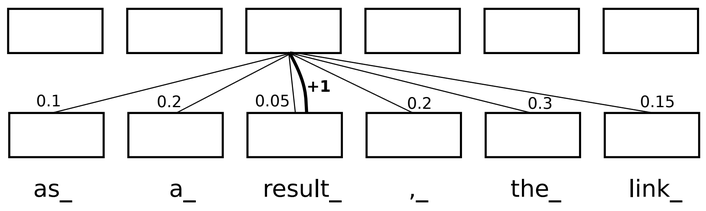
- 6 layers
- 16-head attentions
- 100k wordpieces
-
- English to Czech
- English to German
- English to French
- English to Finnish

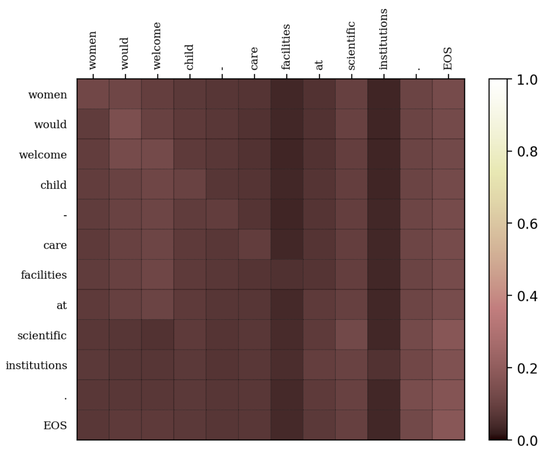


Aggregation over Layers

Due to residual connection, on each the position, one half of information is copied from the previous layer.

We aggregate the attentions through layers to get distribution of the source wordpieces for each position in each layer





We found that on the 6th layer, the distributions over the source wordpieces influencing particular positions are very flat.

We therefore do not work with the aggregated attentions in further experiments.

Preliminary evaluation

Convert manually annotated dependency trees (UD) to unlabelled phrase trees.
Compute the ratio of valid (non-crossing) phrases
- precision: check output phrases against manual
- recall: check manual phrases against output
  (high: manual trees very flat)

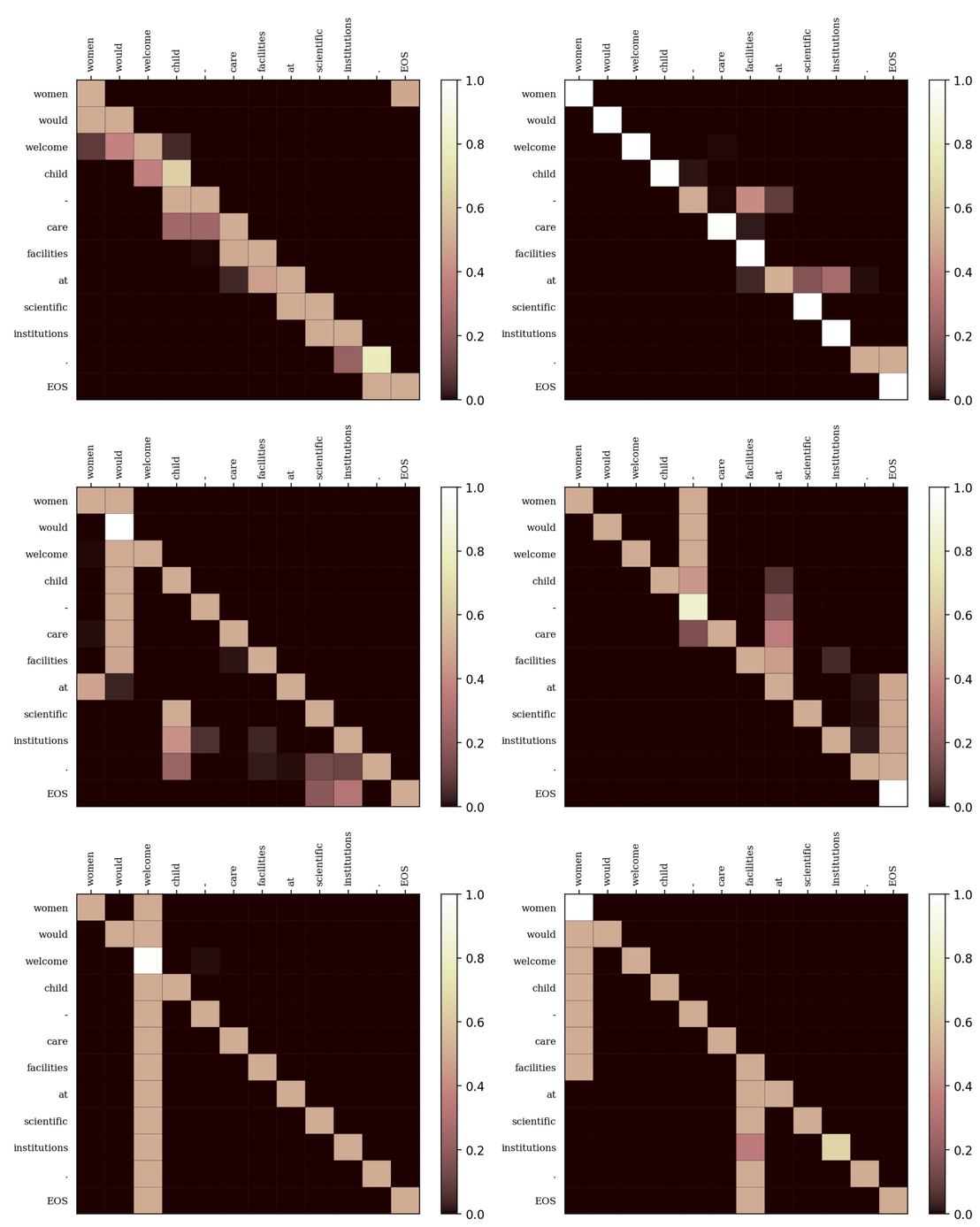
Highest scoring baseline: right-aligned binary tree

Results (on 1st 100 sentences from English to Finnish translation)

Baseline: P=50.74%, R=81.60%
Extracted: P=50.99%, R=83.43%

1. Translate the sentence (e.g. from English to German)
Women would welcome child-care facilities at scientific institutions.
---->
Frauen würden Kinderbetreuungs@@ einrichtungen in wissenschaftlichen Institutionen begrüßen.

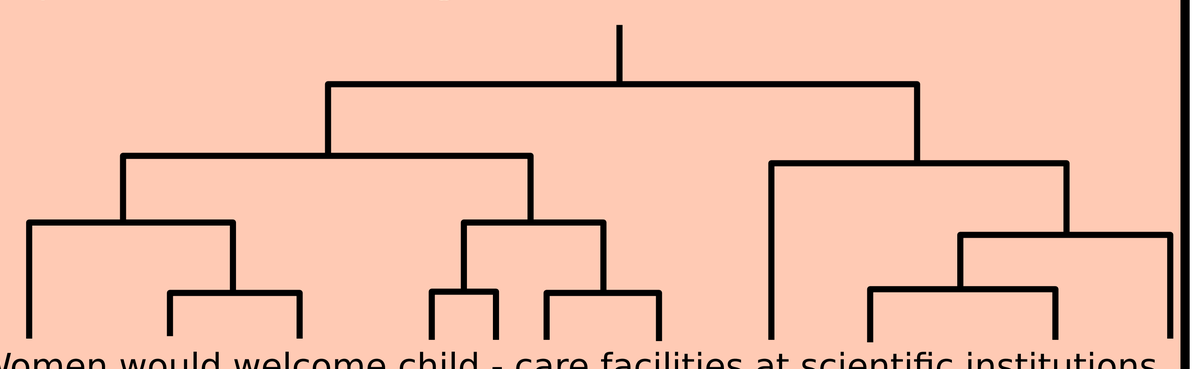
2. Extract self-attention weights
16 heads x 6 layers = 64 matrices (only six of them are shown)



3. Compute score for each possible phrase

We collect continuous sequences from all attention matrices
Score of the phrase = maximal sum of such sequence

4. Use CKY algorithm to parse the sentence
We find the phrase-tree with best scored phrases using dynamic programming.
Probability of joining two phrases and probability of the two respective subtrees is weighted 1:1



Analysis of self-attentions

We found that there are heads that...
- often look on the previous or the following word (mainly the 1st layer)
- the whole phrase looks on its first (or last) word
- if there are equal words in one sentences, they look on each other
- majority of words look to the end of the sentence

ITAT, Krompachy, Slovakia, September 22, 2018

This work has been supported by the grant 18-02196S of the Czech Science Foundation