David Mareček, Rudolf Rosa
marecek@ufal.mff.cuni.cz, rosa@ufal.mff.cuni.cz

From Balustrades to Pierre Vinken:
Looking for Syntax
in Transformer Self-Attention

Charles University, Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
BlackboxNLP Workshop, Firenze, 1 August 2019
huge areas covering thousands of hectares of vineyards have been burned; this means that the wine-growers have suffered loss and that their plants have been damaged.
Observation
Observation

- Common pattern in Transformer NMT self-attention heads
Observation

- Common pattern in Transformer NMT self-attention heads

huge areas covering thousands of hectares of vineyards have been burned; this means that the vinegrowers have suffered loss and that their plants have been damaged.
Observation

- Common pattern in Transformer NMT self-attention heads

huge areas covering thousands of hectares of vineyards have been burned; this means that the winegrowers have suffered loss and that their plants have been damaged.

(d) layer 4 head 13
Observation

- Common pattern in Transformer NMT self-attention heads

huge areas covering thousands of hectares of vineyards have been burned; this means that the vinegrowers have suffered loss and that their plants have been damaged.
- **Common pattern in Transformer NMT self-attention heads**
Observation

- Common pattern in Transformer NMT self-attention heads

huge areas covering thousands of hectares of vineyards have been burned; this means that the winegrowers have suffered loss and that their plants have been damaged.
- Common pattern in Transformer NMT self-attention heads

huge areas covering thousands of hectares of vineyards have been burned; this means that the wine-growers have suffered loss and that their plants have been damaged.
Observation

- Common pattern in Transformer NMT self-attention heads

 huge areas covering thousands of hectares of vineyards have been burned; this means that the winegrowers have suffered loss and that their plants have been damaged.
Observation

- Common pattern in Transformer NMT self-attention heads
Observation

- Common pattern in Transformer NMT self-attention heads
  - “balusters”
Observation

- Common pattern in Transformer NMT self-attention heads
  - “balusters”
- Resemble syntactic phrases
Observation

- Common pattern in Transformer NMT self-attention heads
  - “balusters”
- Resemble syntactic phrases

(d)
layer 4
head 13

huge areas covering thousands of hectares of vineyards have been burned; this means that the vine-growers have suffered loss and that their plants have been damaged.
Observation

- Common pattern in Transformer NMT self-attention heads
  - “balusters”
  - Resemble syntactic phrases
    - To what extent?
      → That’s our research question!

(d) layer 4 head 13

huge areas covering thousands of hectares of vineyards have been burned; this means that the vinegrowers have suffered loss and that their plants have been damaged.

Mareček & Rosa: From Balustrades to Pierre Vinken: Looking for Syntax in Transformer Self-Attentions
Approach
Approach

1. Balusters → phrase candidates
1. Balusters → phrase candidates
2. Phrase candidates → constituency tree
   - Linguistically uninformed algorithm
Approach

1. Balusters $\rightarrow$ phrase candidates
2. Phrase candidates $\rightarrow$ constituency tree
   - Linguistically uninformed algorithm
3. Compare to standard syntactic trees
1. Balusters → phrase candidates

2. Phrase candidates → constituency tree
   - Linguistically uninformed algorithm

3. Compare to standard syntactic trees: ~40%; baseline ~30%
Experiment setup

- Balusters: Transformer NMT system
  - Encoder: 6 layers x 16 heads
Experiment setup

- **Balusters: Transformer NMT system**
  - Encoder: 6 layers x 16 heads
  - Europarl: French ↔ English, German ↔ English, French ↔ German
Experiment setup

- **Balusters: Transformer NMT system**
  - Encoder: 6 layers x 16 heads
  - Europarl: French ↔ English, German ↔ English, French ↔ German

- **Standard syntactic trees: Stanford parser**
  - Penn Treebank, French Treebank, Negra Corpus
  - Only for evaluation
huge areas covering thousands of hectares of vineyards have been burned; this means that the vinegrowers have suffered loss and that their plants have been damaged.

Mareček & Rosa: From Balustrades to Pierre Vinken: Looking for Syntax in Transformer Self-Attentions
Diagonals (especially 1\textsuperscript{st} layer)
Attend to end, mixed, scattered...
Phrase candidates

- All balusters of length $\geq 2$ from **all** heads
  - Subselecting only some of the heads → see the paper!
Phrase candidates

- All balusters of length $\geq 2$ from all heads
  - Subselecting only some of the heads $\rightarrow$ see the paper!
- Phrase score
  - Average attention weight
  - Sum over all heads
  - Equalize over different phrase lengths
- All balusters of length $\geq 2$ from all heads
  - Subselecting only some of the heads → see the paper!
- Phrase score
  - Average attention weight
  - Sum over all heads
  - Equalize over different phrase lengths

huge areas covering thousands of hectares of vineyards have been burned; this means that the winegrowers have suffered loss and that their plants have been damaged.
Phrase candidates → constituency tree
Phrase candidates $\rightarrow$ constituency tree

- Binary constituency tree

```
   X
  / \  /
 X   X a b c
```

Mareček & Rosa: From Balustrades to Pierre Vinken: Looking for Syntax in Transformer Self-Attentions
Phrase candidates → constituency tree

- Binary constituency tree
- Tree score = sum of phrase scores

\[
s(T) = s(ab) + s(abc)
\]
Phrase candidates → constituency tree

- Binary constituency tree
- Tree score = sum of phrase scores
- CKY algorithm
  - Finds tree (set of phrases) with maximal score

\[ s(T) = s(ab) + s(abc) \]
Results

huge areas covering thousands of hectares of vineyards have been burned; this means that the vine growers have suffered loss and that their plants have been damaged.
Results

huge areas
covering
thousands of hectares
of vineyards
have been burned

; this means that
the vine growers have suffered loss and that

these huge hectares of vineyards have been burned.
Comparison to standard syntactic trees

- EN->DE
- EN->FR
- DE->EN
- DE->FR
- FR->EN
- FR->DE

Base
Our
Summary
Summary

- **Balusters** in Transformer NMT encoder self-attentions
  - Contiguous sequence of output states
  - Attention to the same one input state
Summary

- **Balusters** in Transformer NMT encoder self-attentions
  - Contiguous sequence of output states
  - Attention to the same one input state
- Interpret balusters as **syntactic phrases**
  - Phrase candidate extraction and scoring
- Construct a binary **constituency tree**
  - CKY algorithm

(d) layer 4 head 13

huge areas covering thousands of hectares of vineyards have been burned; this means that the vine-...
Summary

- **Balusters** in Transformer NMT encoder self-attentions
  - Contiguous sequence of output states
  - Attention to the same one input state
- Interpret balusters as **syntactic phrases**
  - Phrase candidate extraction and scoring
- Construct a binary **constituency tree**
  - CKY algorithm
- Compare to **standard syntactic trees**
  - ~40% match; base ~30% match
Thank you for your attention

David Mareček, Rudolf Rosa
marecek@ufal.mff.cuni.cz, rosa@ufal.mff.cuni.cz

From Balustrades to Pierre Vinken:
Looking for Syntax in Transformer Self-Attentions

Charles University, Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

ufal.cz/grants/lsd