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

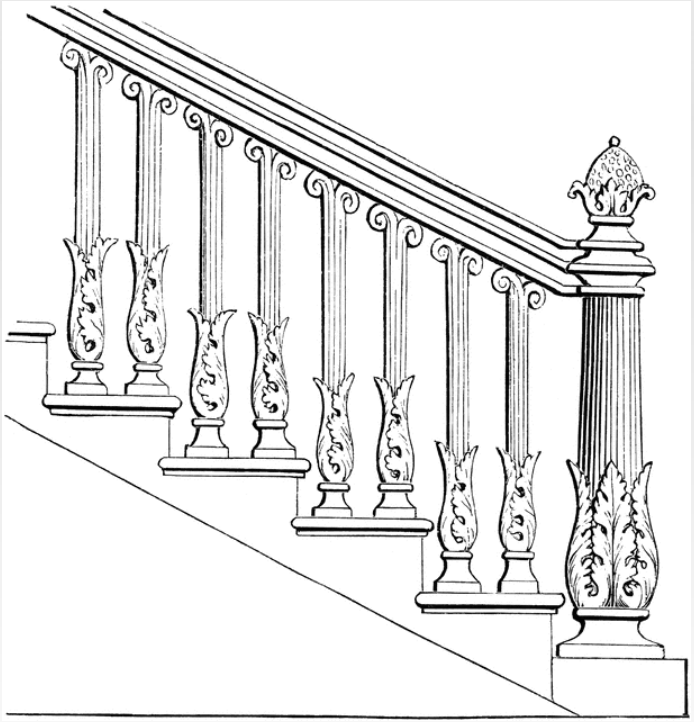
From Balustrades to Pierre Vinken:

Looking for Syntax in Transformer Self-Attentions

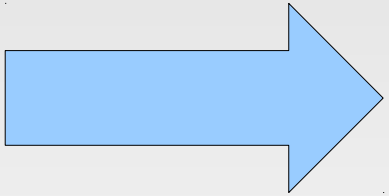
Charles University, Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

BlackboxNLP Workshop, Firenze, 1 August 2019

From balustrades to Pierre Vinken

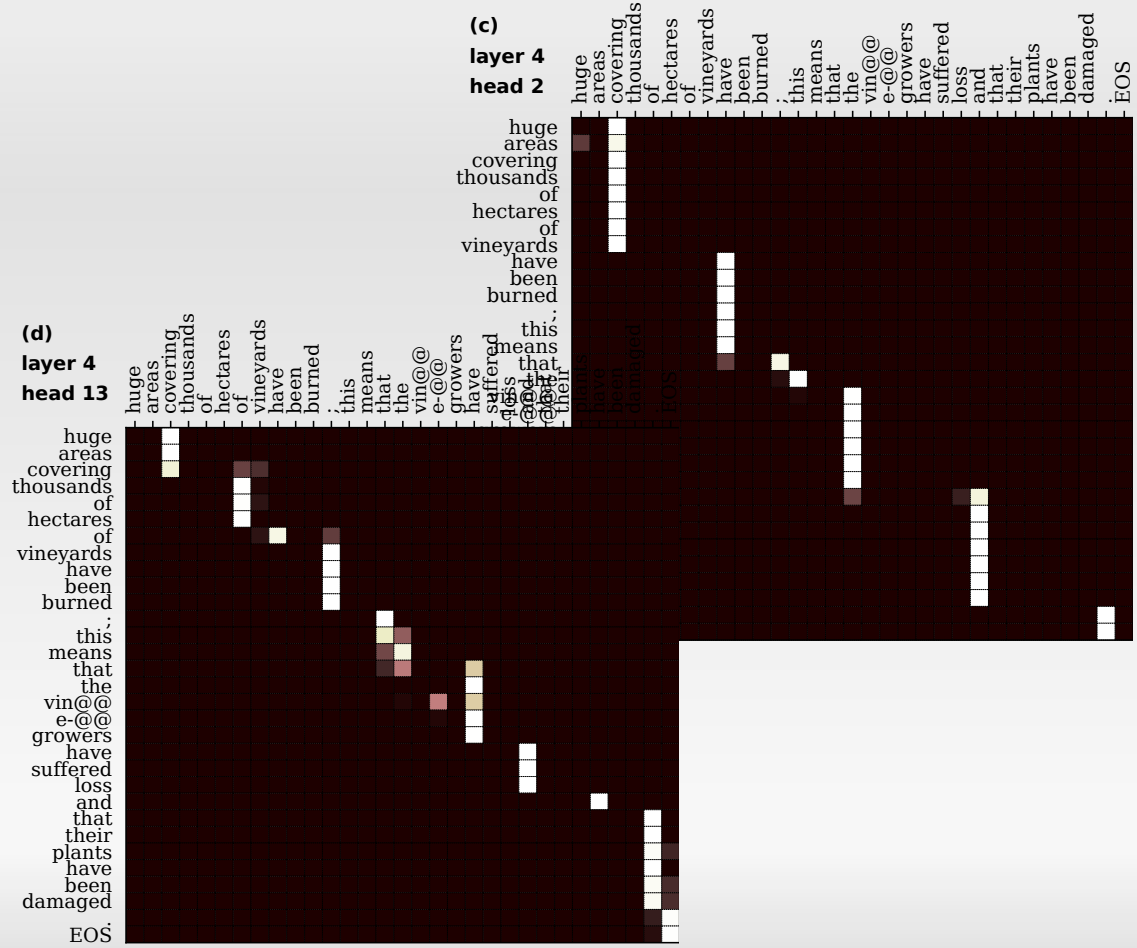


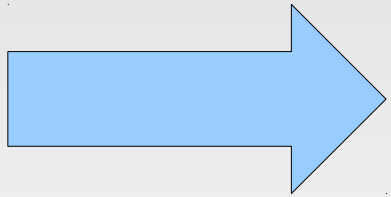
http://clipart-library.com/clipart/28144.htm

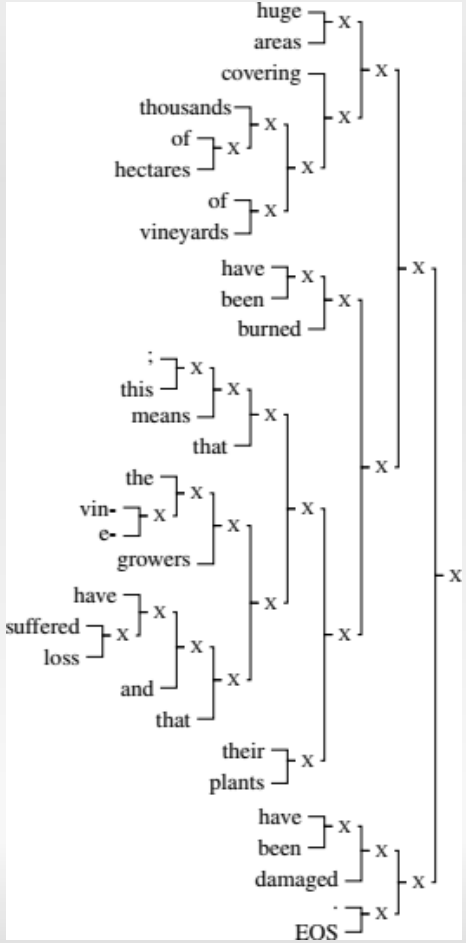


by Jan Hein van Dierendonck

Transformer self-attentions → syntactic trees







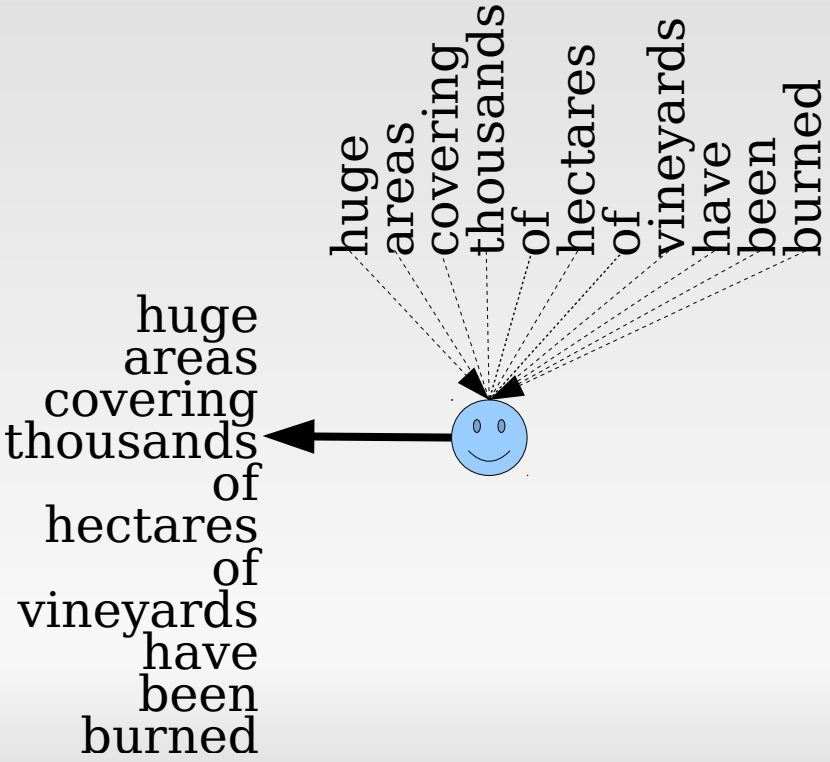
Observation

Observation

- Common pattern in Transformer NMT self-attention heads

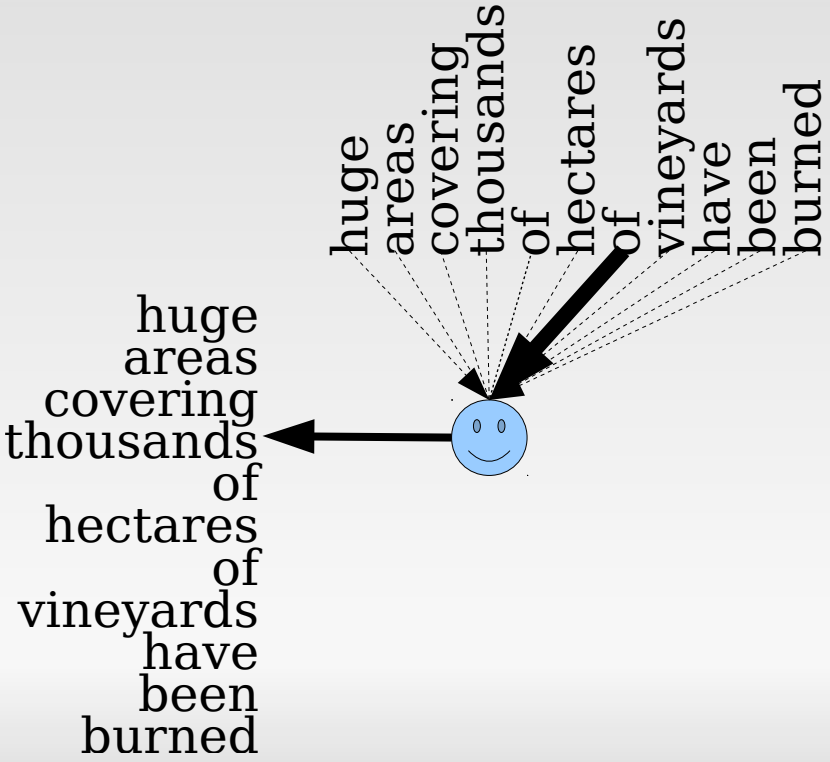
Observation

- Common pattern in Transformer NMT self-attention heads



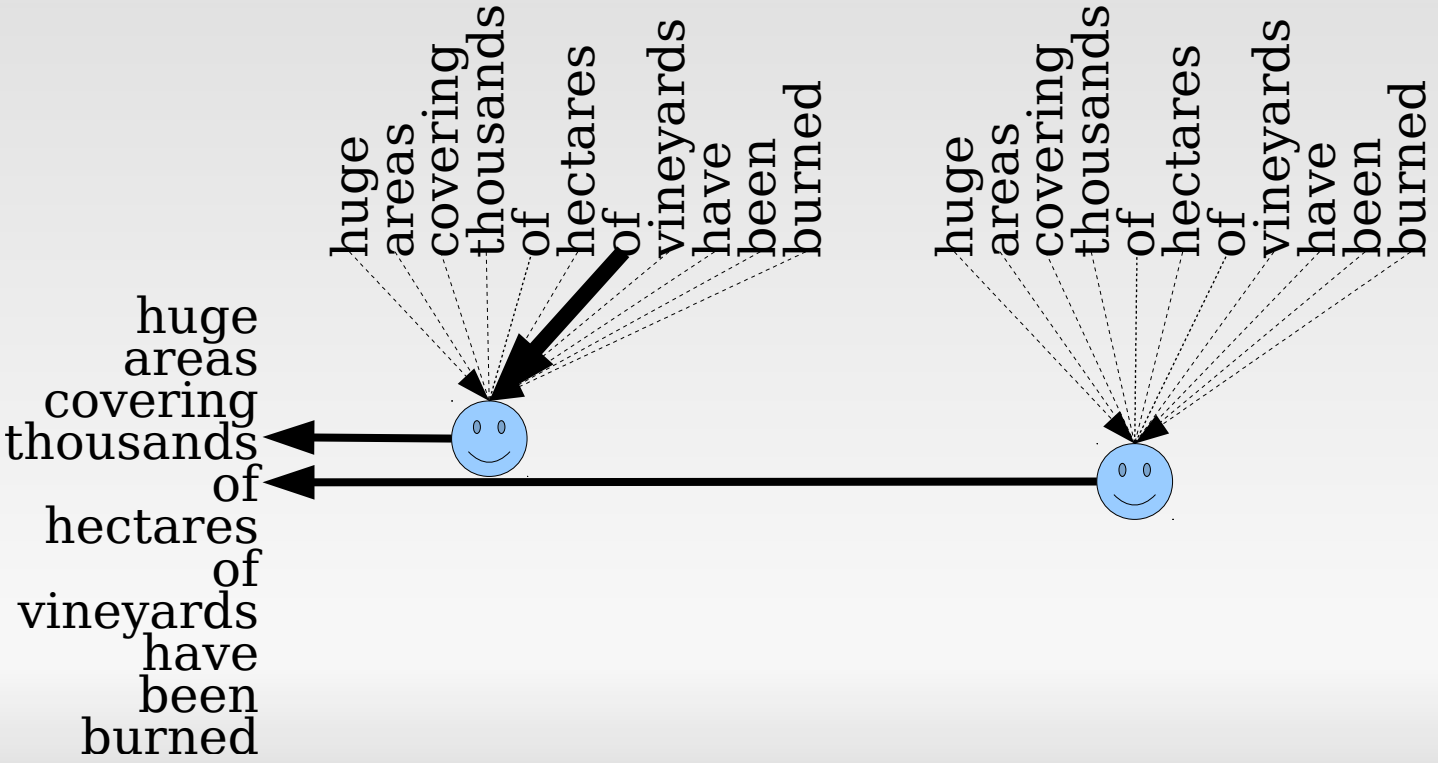
Observation

- Common pattern in Transformer NMT self-attention heads



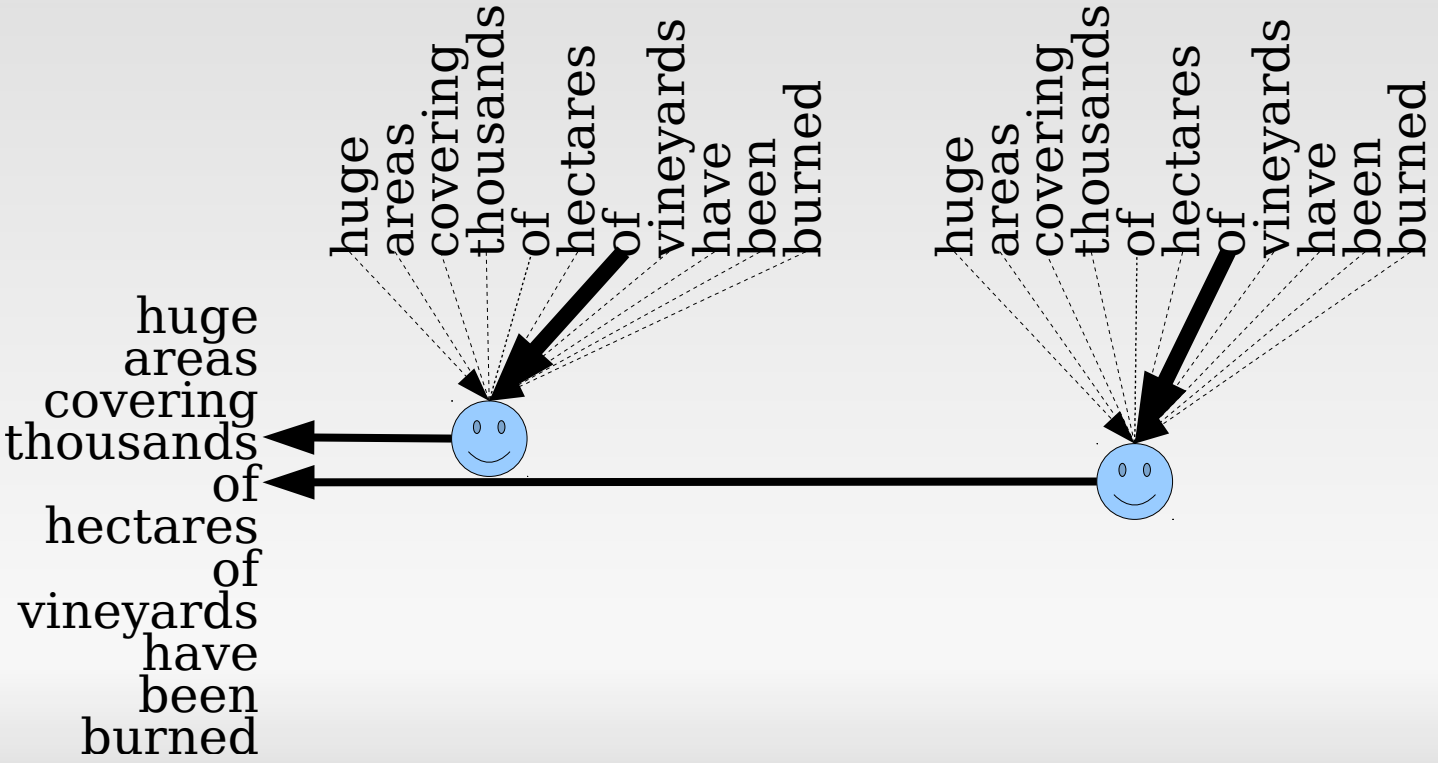
Observation

- Common pattern in Transformer NMT self-attention heads



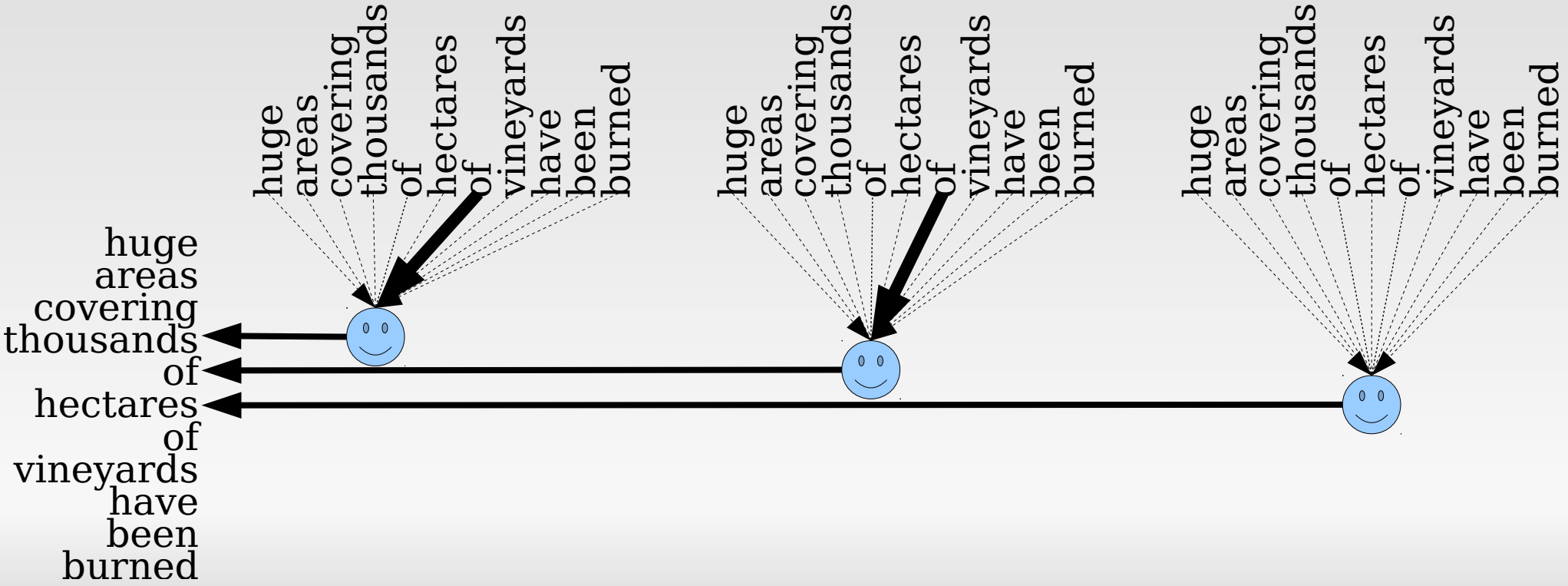
Observation

- Common pattern in Transformer NMT self-attention heads



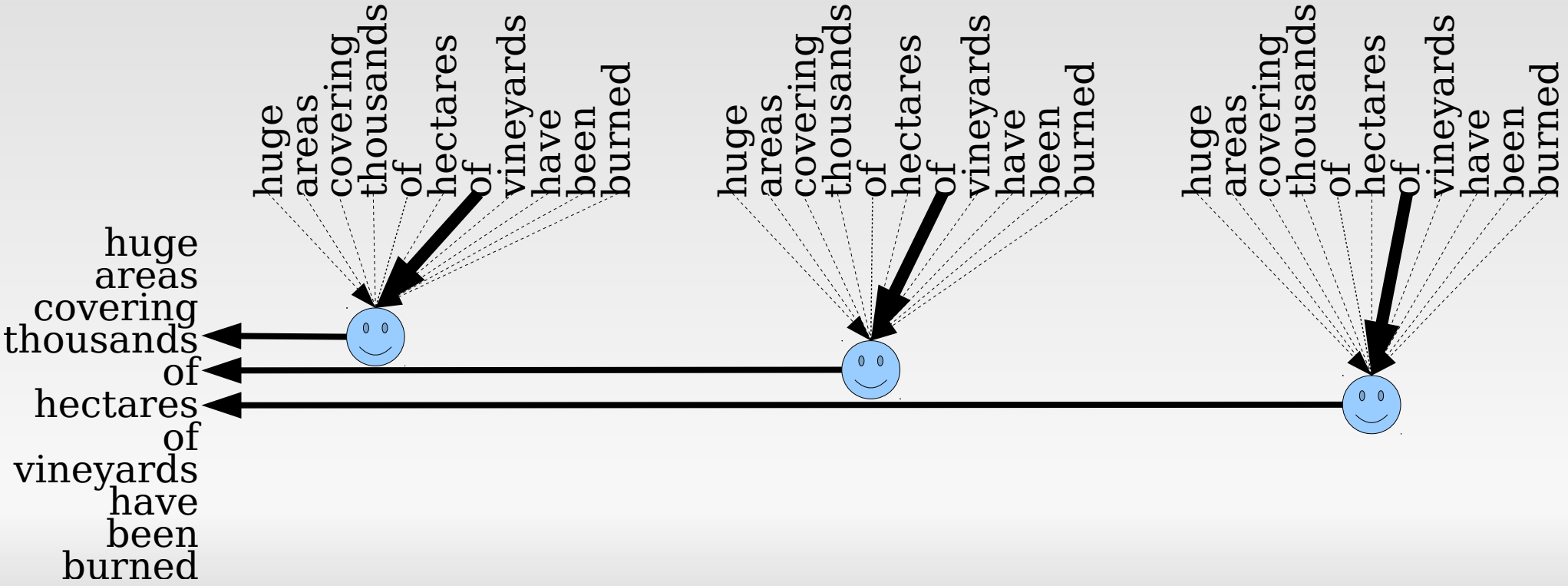
Observation

- Common pattern in Transformer NMT self-attention heads



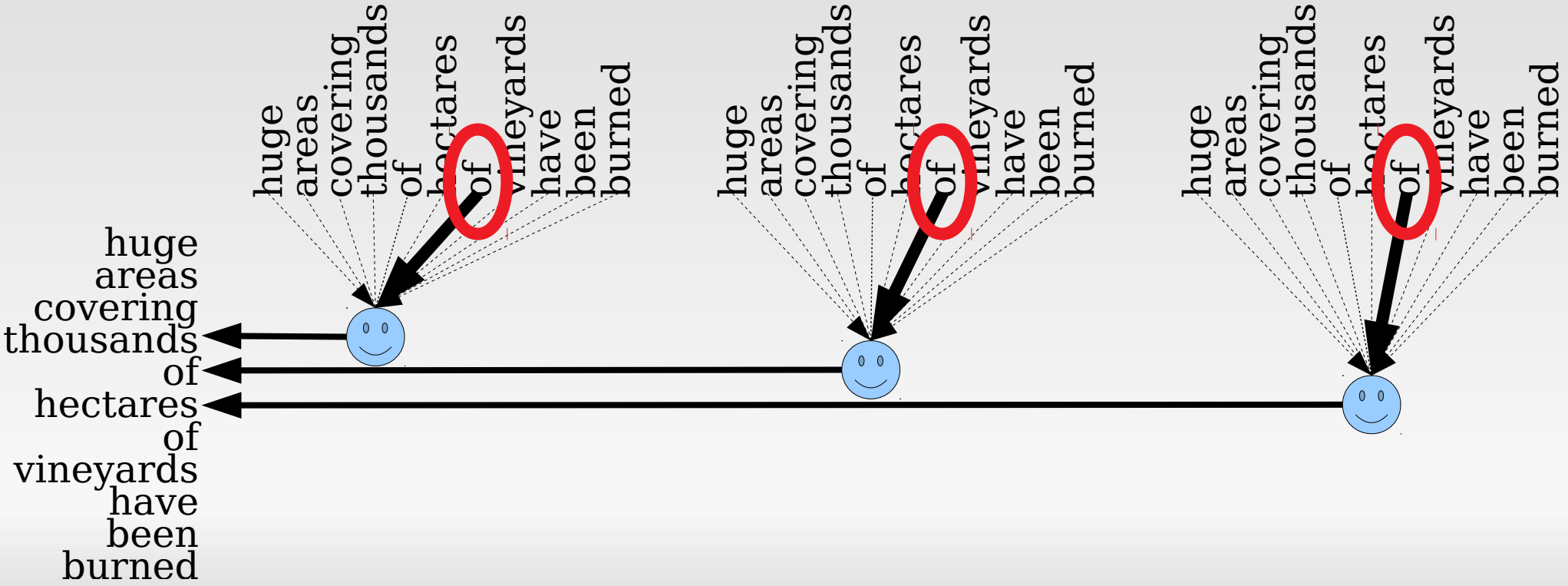
Observation

- Common pattern in Transformer NMT self-attention heads



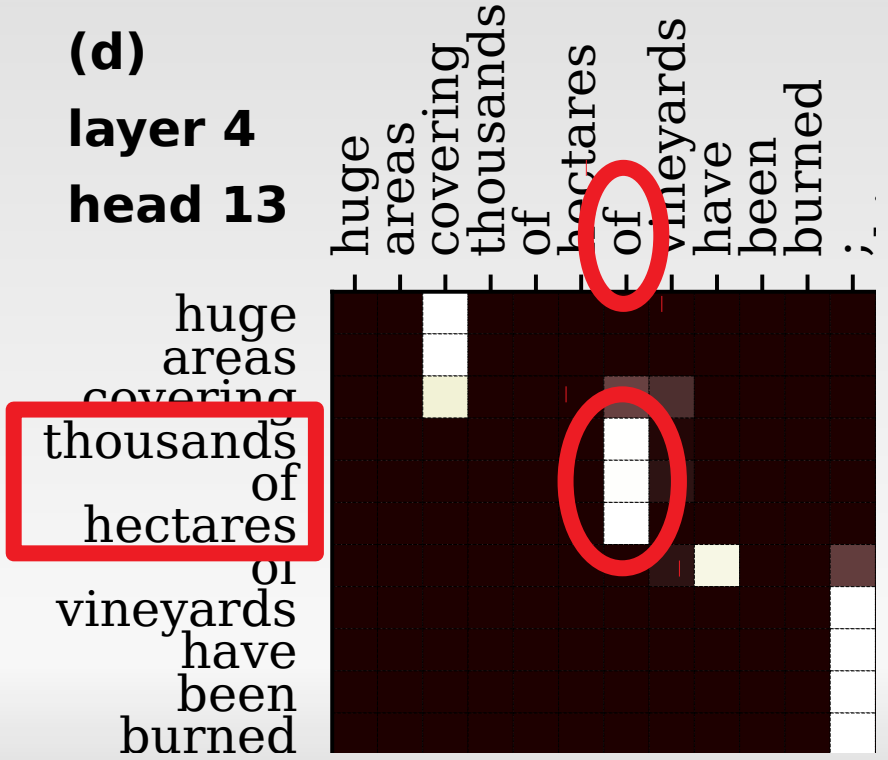
Observation

- Common pattern in Transformer NMT self-attention heads



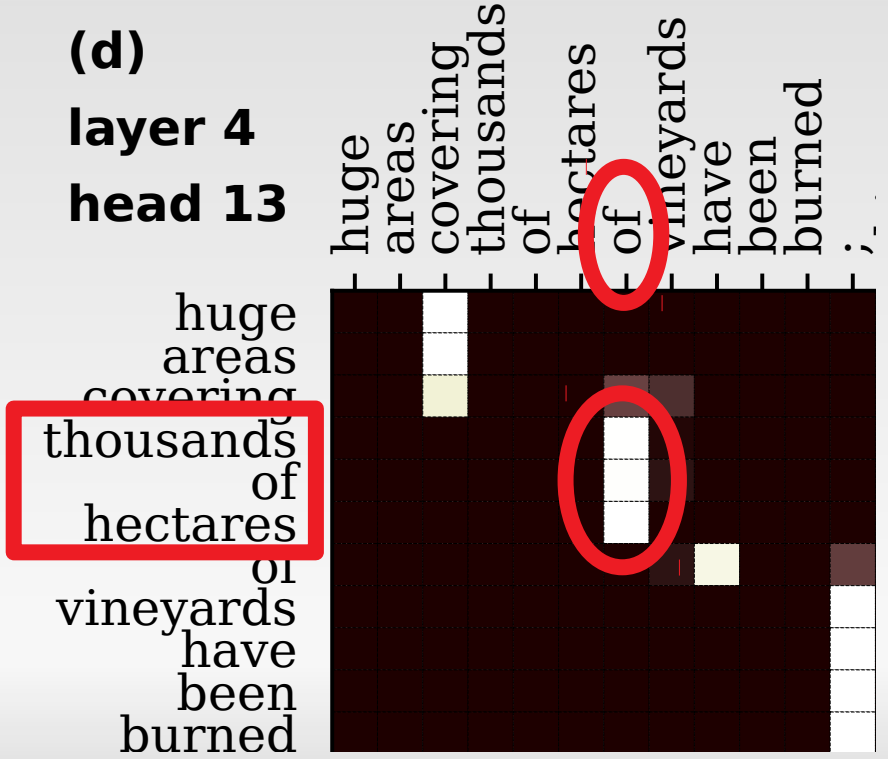
Observation

- Common pattern in Transformer NMT self-attention heads



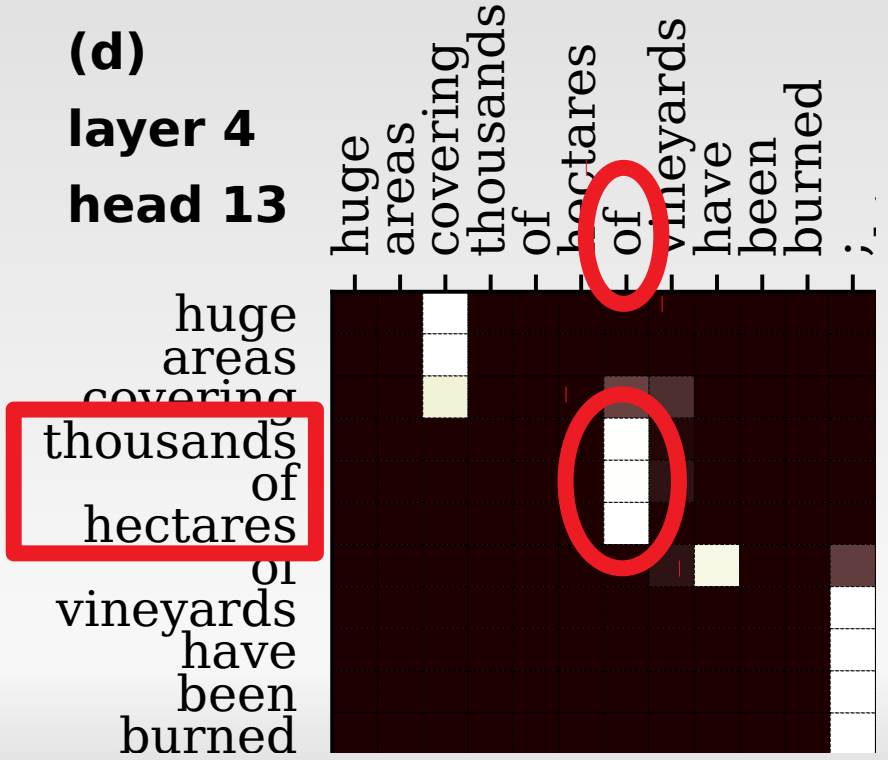
Observation

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



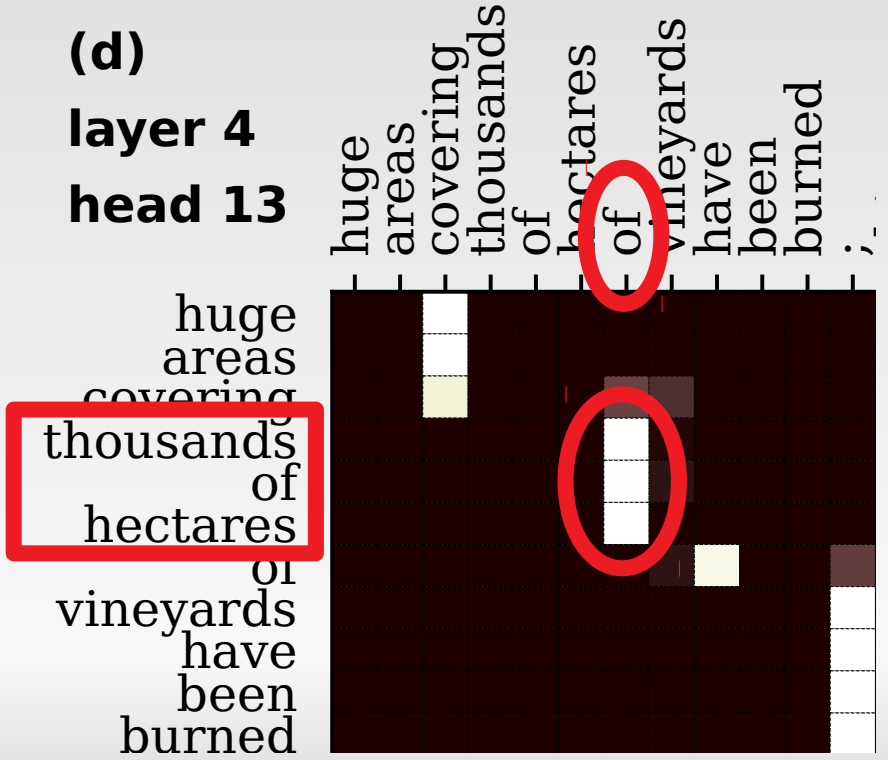
Observation

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



Observation

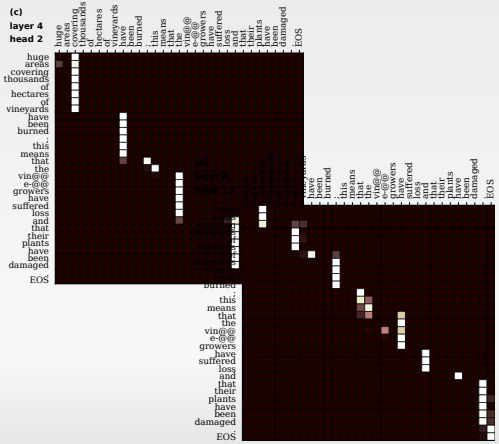
- Common pattern in Transformer NMT self-attention heads
  - “balusters”
- Resemble syntactic phrases
  - To what extent?
    → That’s our research question!



Approach

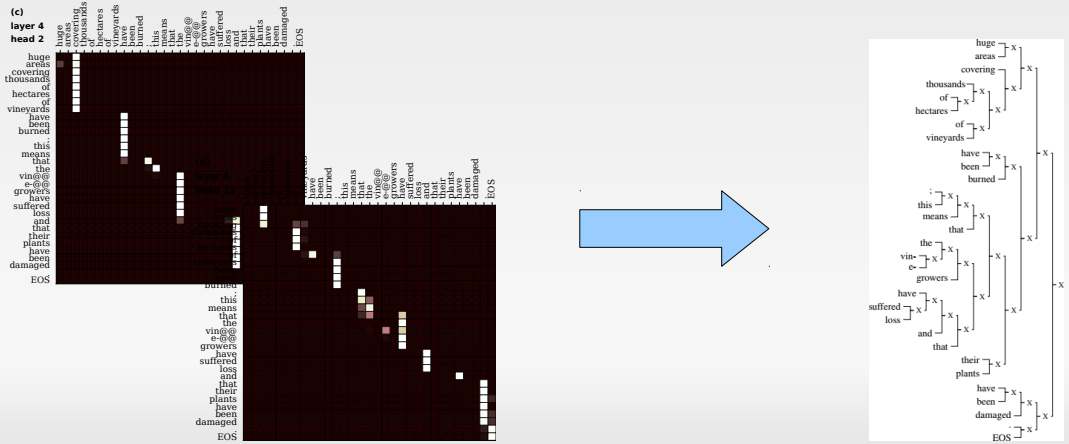
Approach

1. Balusters → phrase candidates



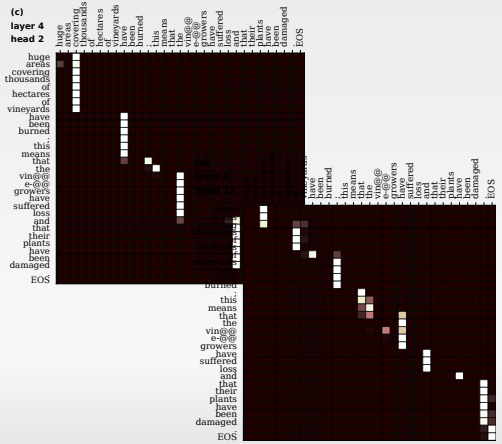
Approach

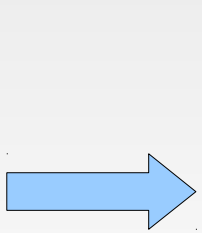
1. Balusters → phrase candidates
2. Phrase candidates → constituency tree
   - Linguistically uninformed algorithm

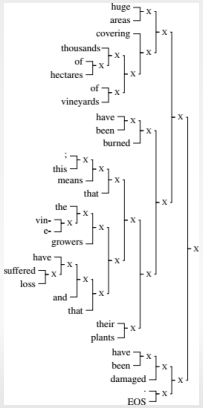


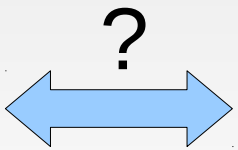
Approach

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



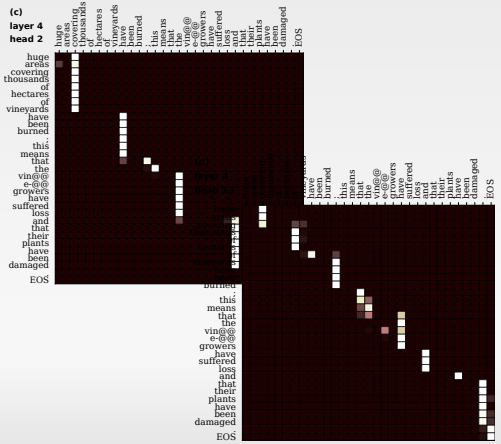


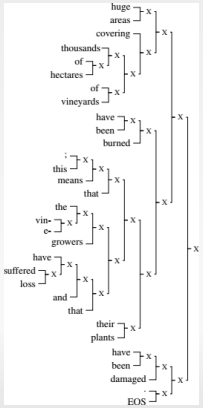


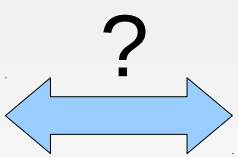


Approach

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

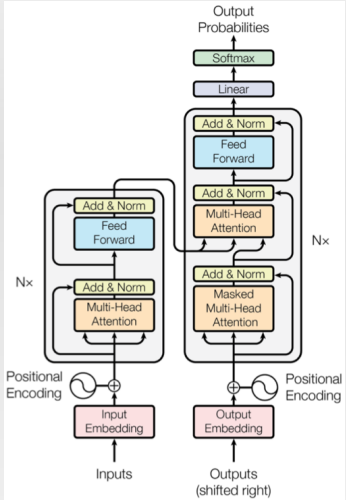


Figure 1: The Transformer - model architecture.

Experiment setup

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

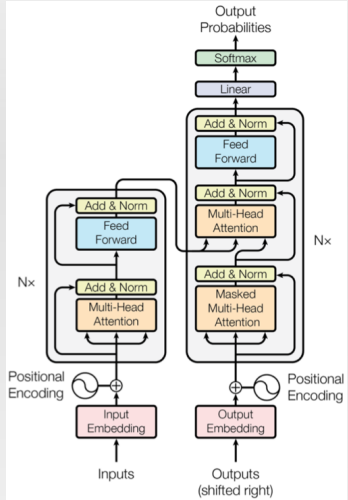
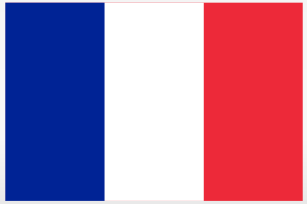


Figure 1: The Transformer - model architecture.

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



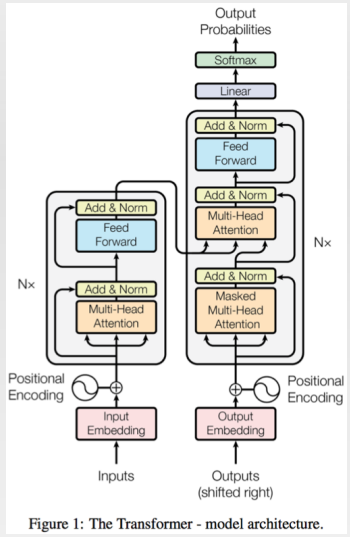
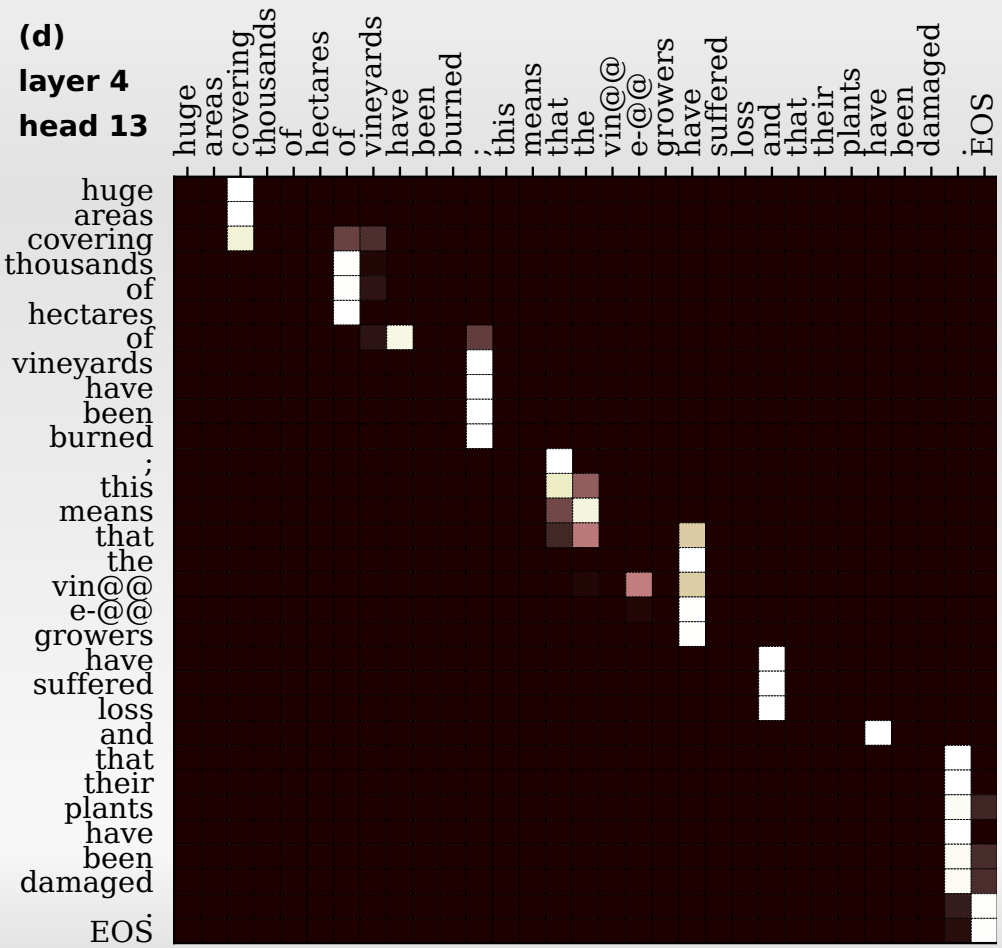
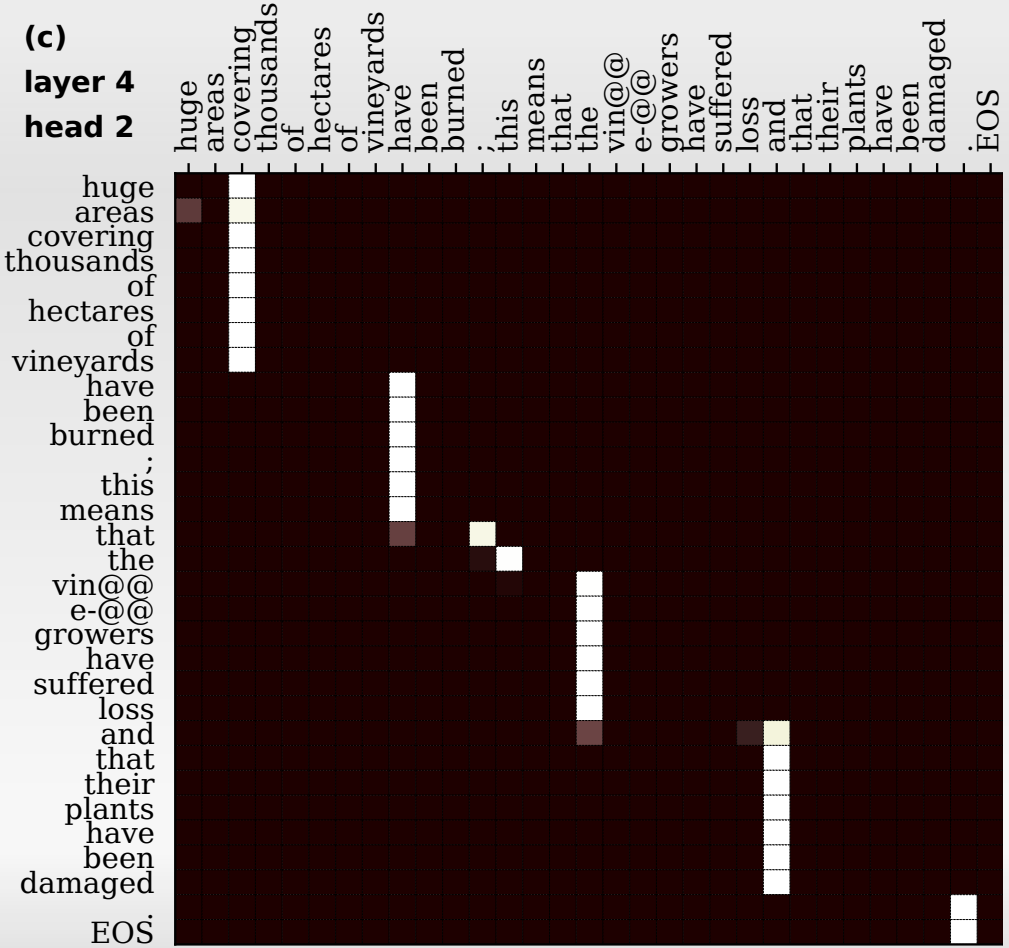


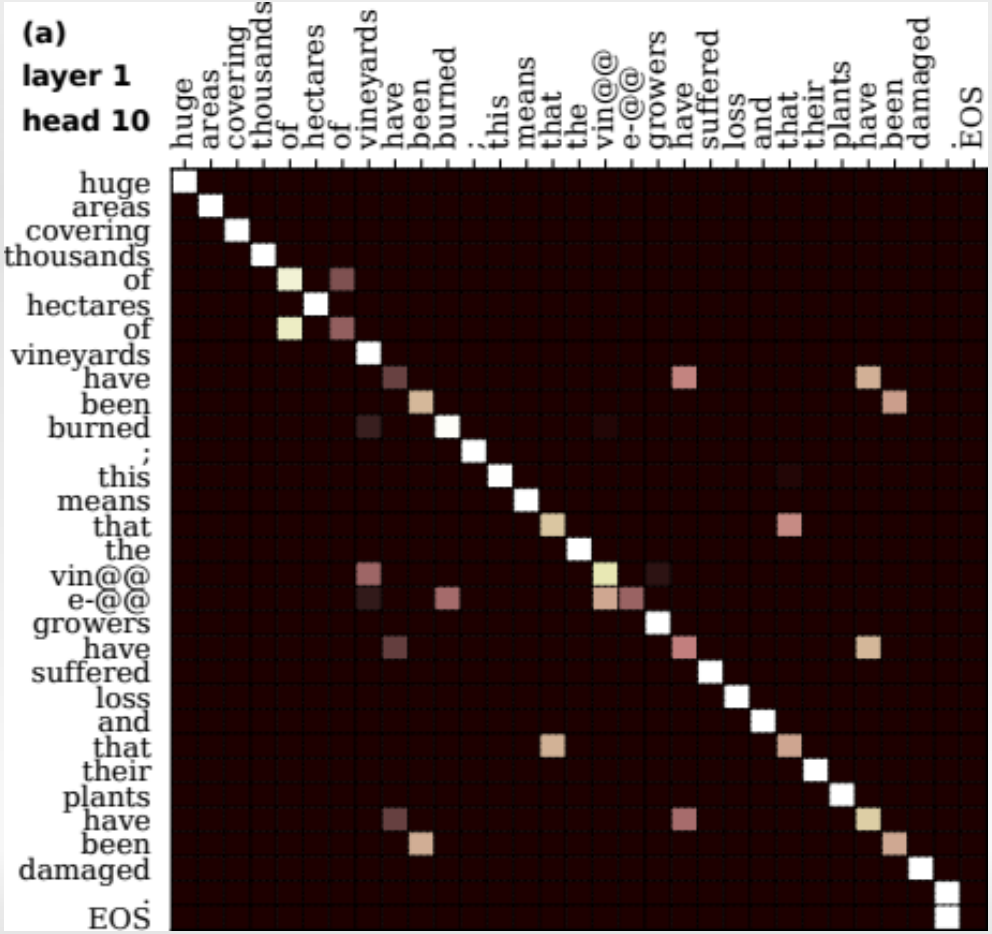
Figure 1: The Transformer - model architecture.

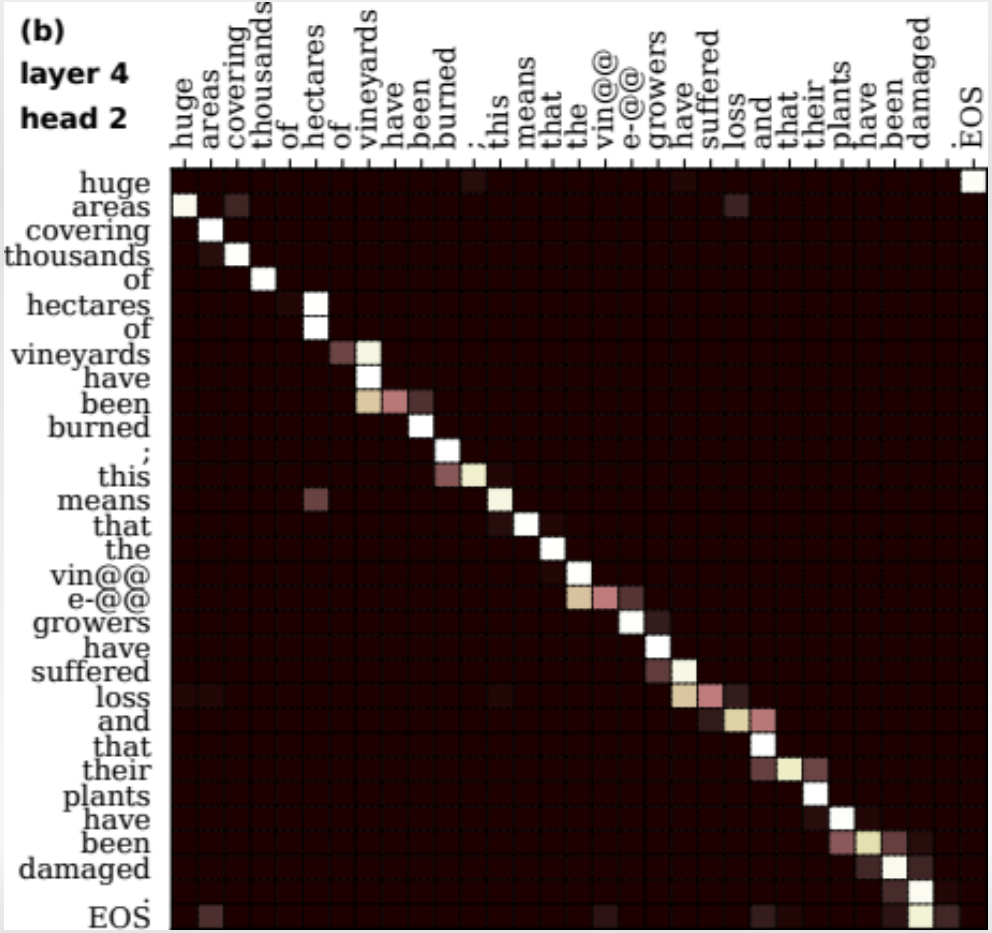
Balustrades (~70% of the attention heads)



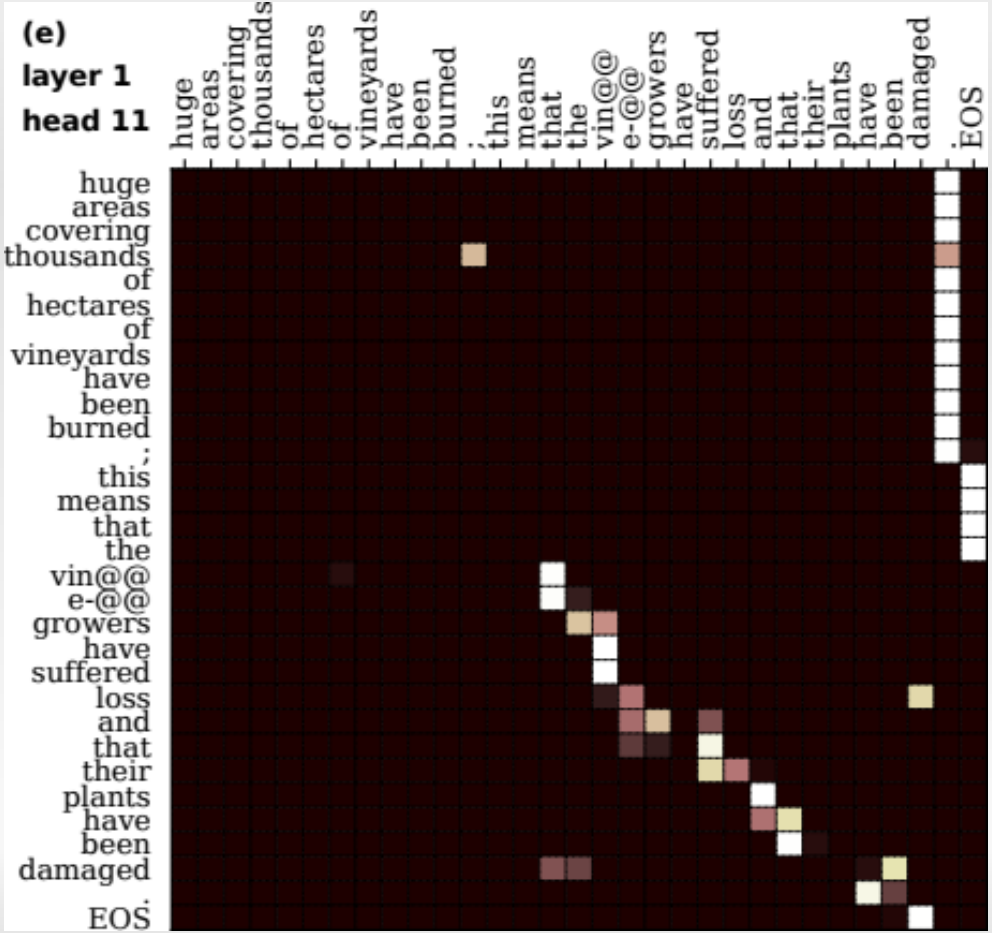


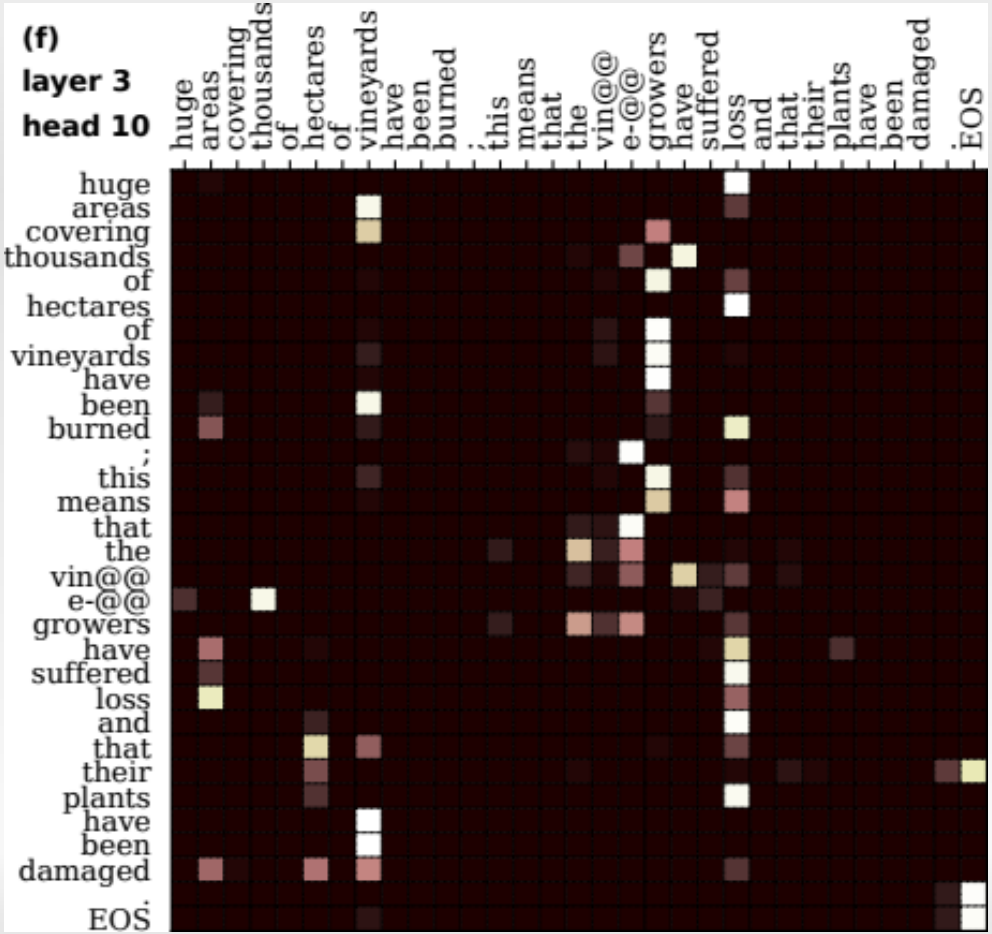
Diagonals (especially 1st layer)





Attend to end, mixed, scattered…





Phrase candidates

- All balusters of length ≥ 2 from **all** heads
  - Subselecting only some of the heads: see the paper!

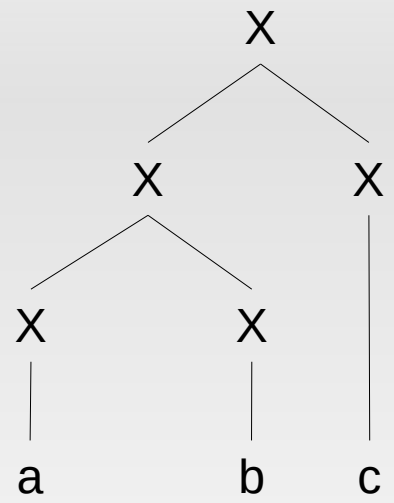
Phrase candidates

- All balusters of length ≥ 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

Phrase candidates → constituency tree

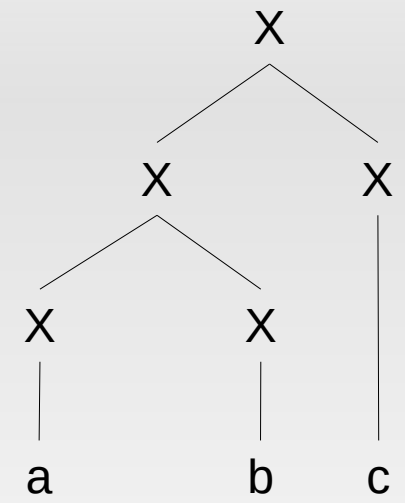
Phrase candidates → constituency tree

- Binary constituency tree



Phrase candidates → constituency tree

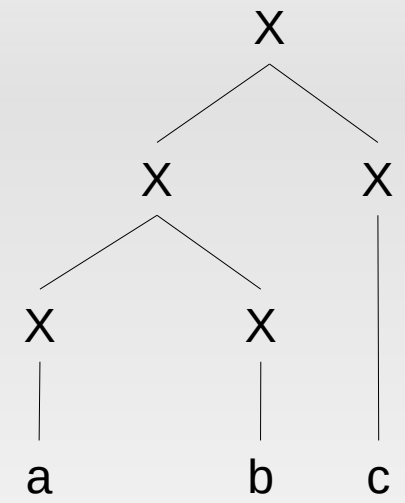
- Binary constituency tree
- Tree score = sum of phrase scores



$s(T) = s(ab) + s(abc)$

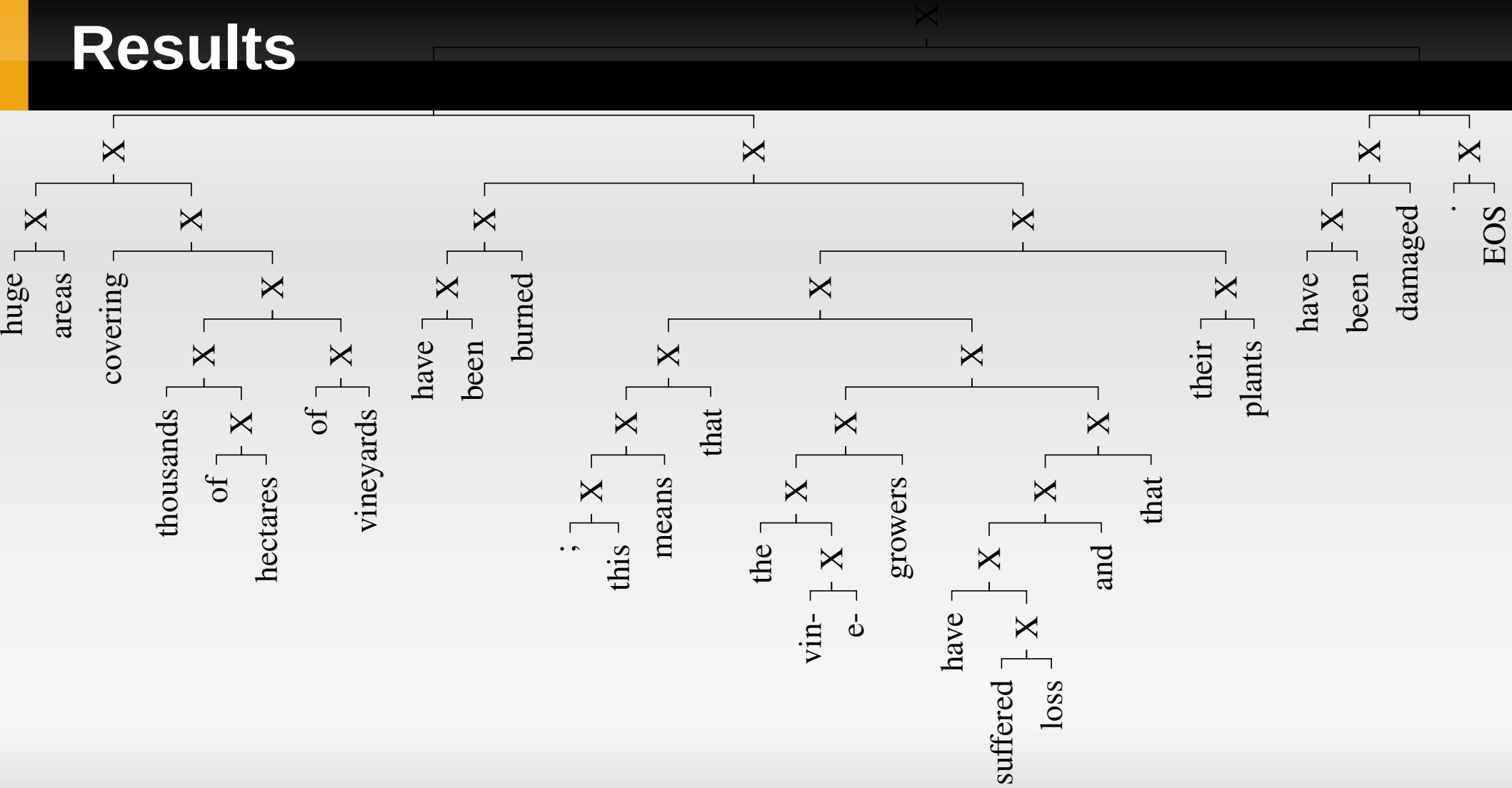
Phrase candidates → constituency tree

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

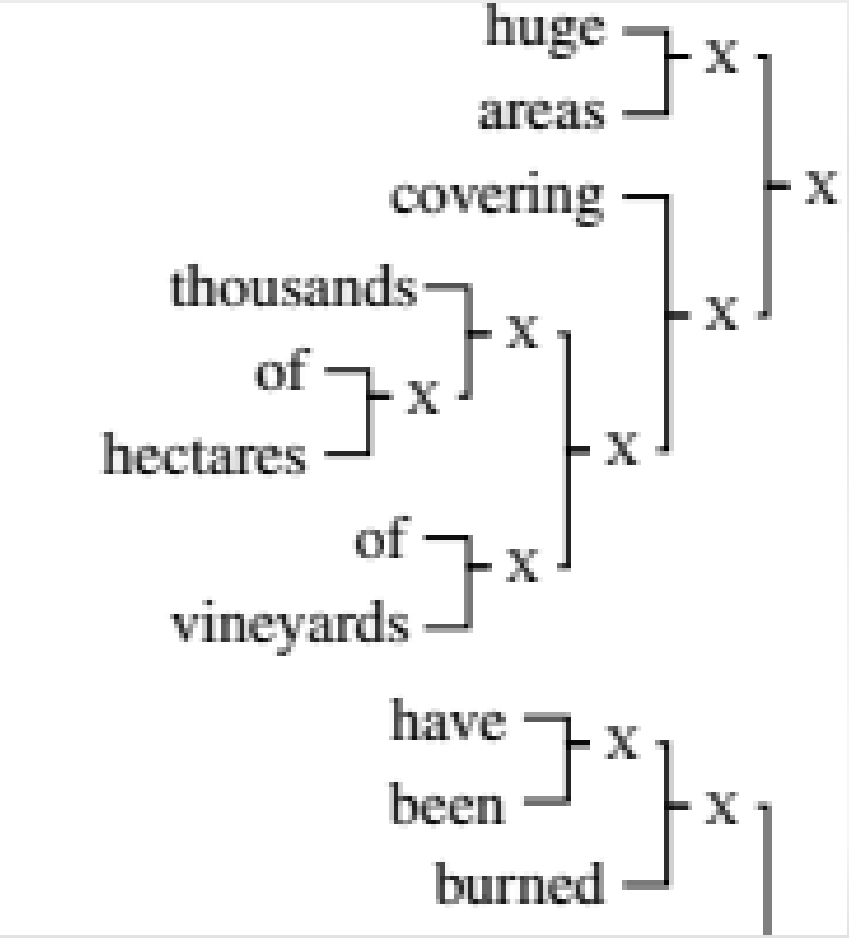


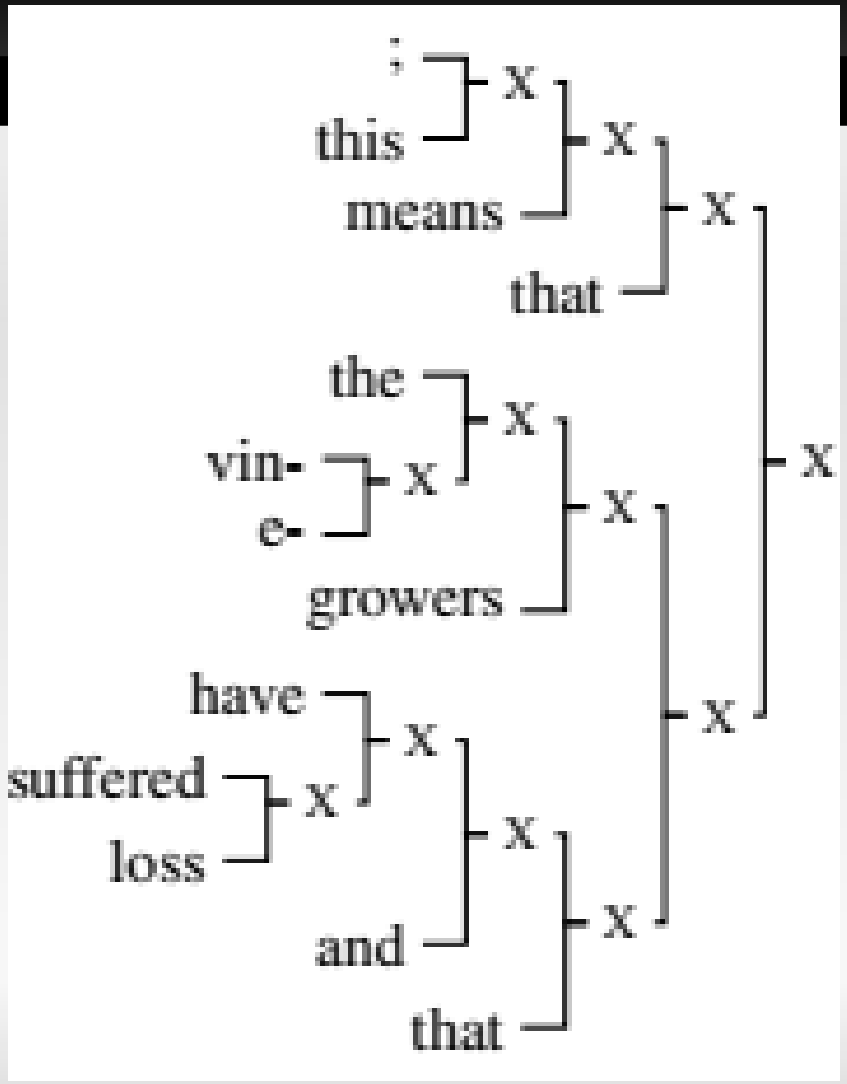
s(T) = s(ab) + s(abc)

Results


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

Results


huge
areas
x
covering
x
x
thousands
of
x
x
hectares
x
of
x
vineyards
have
been
x
x
burned


;
this
x
means
x
that
x
the
vin-
x
e-
x
growers
x
have
suffered
x
loss
x
and
x
that
x
x

Results

English

system	precision	recall	F1 score
rbal	30.1%	24.3%	26.8%
lbal	27.8%	20.8%	23.8%
rand.init	25.1%	20.0%	22.3%
en → de	35.4%	30.6%	32.8%
en → fr	35.4%	30.2%	32.6%

German

system	precision	recall	F1 score
rbal	39.1%	31.3%	34.8%
lbal	38.1%	27.6%	32.0%
rand.init	33.7%	25.9%	29.3%
de → en	46.1%	39.6%	42.6%
de → fr	46.7%	40.9%	43.6%

French

system	precision	recall	F1 score
rbal	34.3%	28.7%	31.3%
lbal	32.5%	25.4%	28.5%
rand.init	26.1%	24.4%	25.3%
fr → en	44.4%	39.7%	41.9%
fr → de	46.9%	41.7%	44.2%

Table 2: Scores of baseline trees and our extracted trees using all attention heads, evaluated against standard syntactic parse trees.

Summary

- Transformer NMT encoder self-attentions
  - diagonals, shifted diagonals, scattered attention…
  - balustrades: can be interpreted as phrases
- Linguistically uninformed syntax extraction
  - baluster → phrase, attention weight → phrase score
  - binary constituency parsing using CKY
  - no training, no hyperparameters, using all heads
    - see the paper for subselecting only some heads
- Resulting structures are quite syntactically sane
  - F1 score 6 – 13 points above baseline (30% → 40%)

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