# Subword pooling

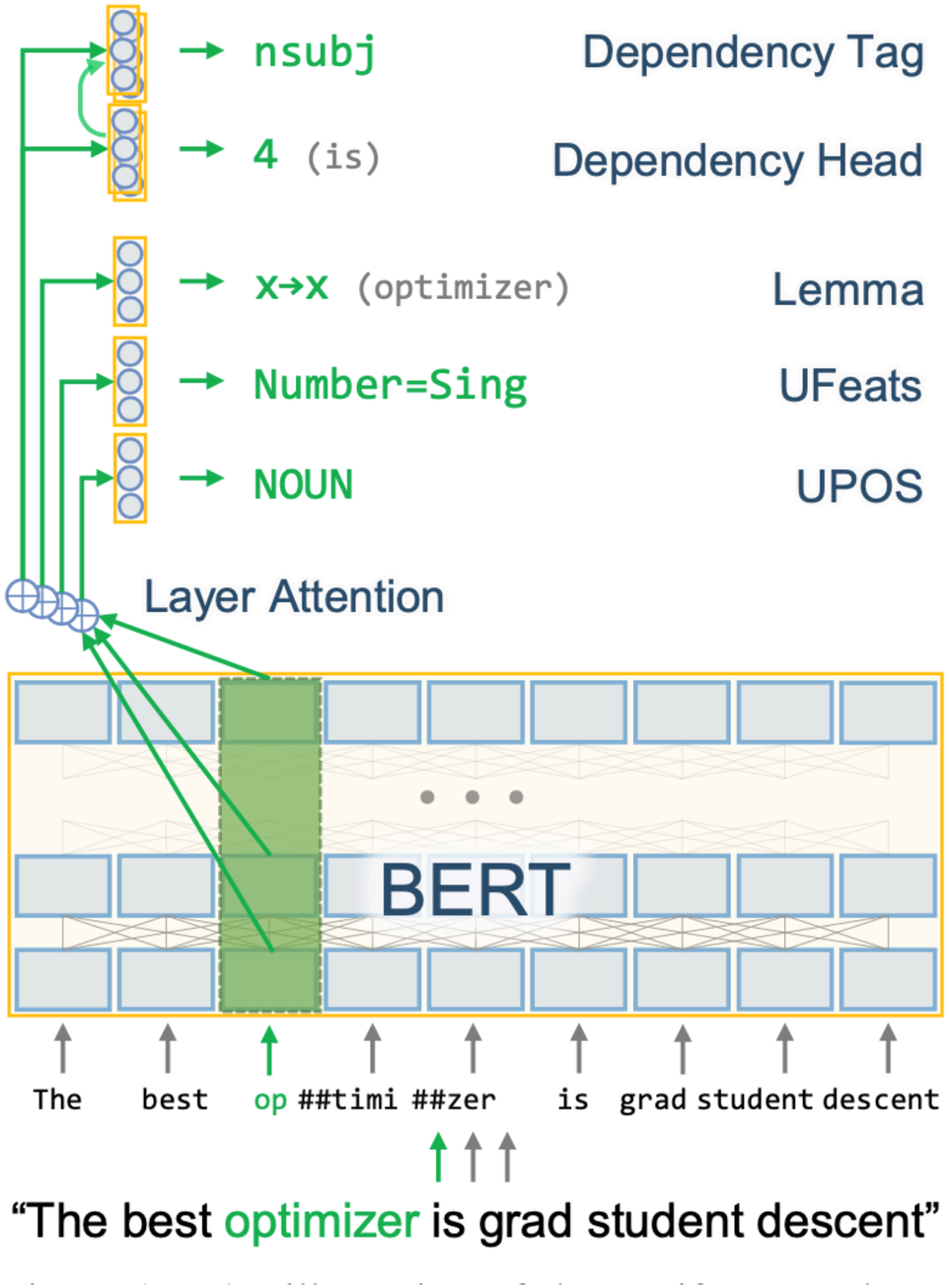
## UniDive WG3 meeting
## Istanbul

**Tanja Samardžić 08.09.2023**

# Usual subword pooling

Kondratyuk and Straka

EMNLP 2019



# Alternative subword pooling

Acs et al.

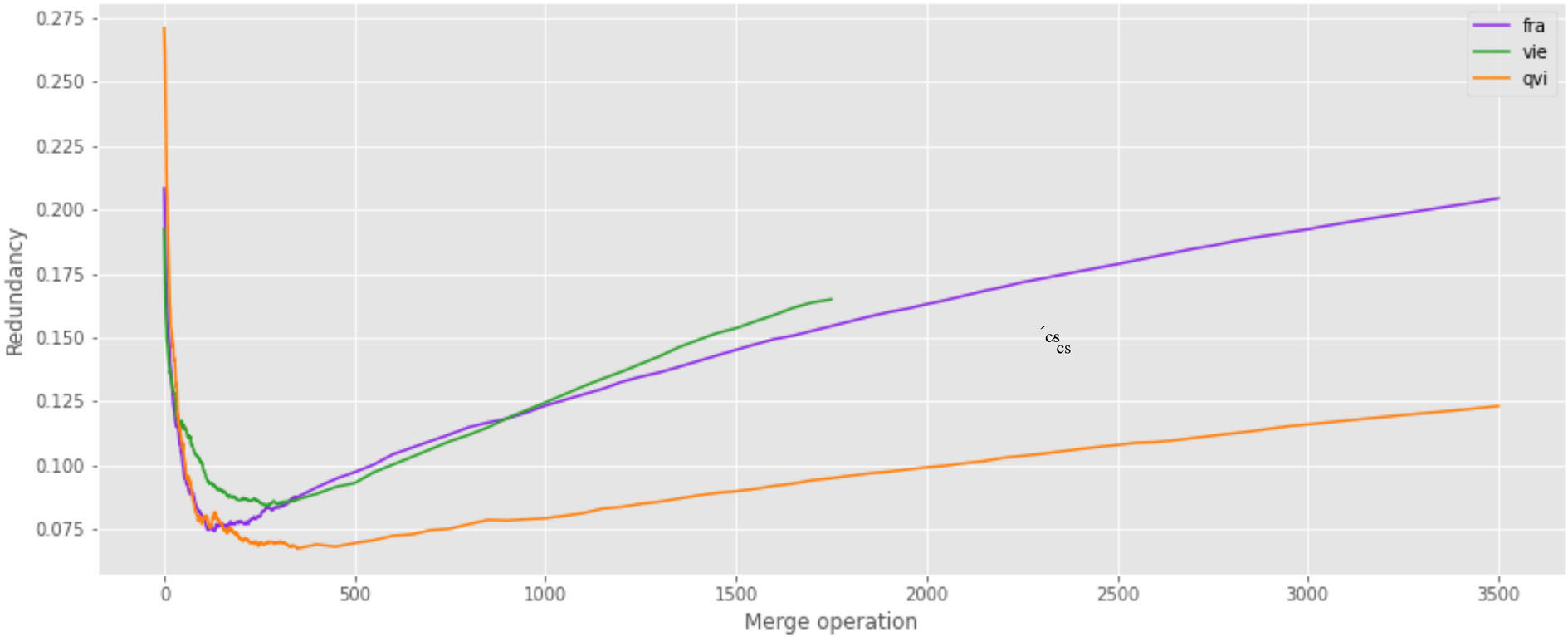
EACL 2021

| Method | Explanation | Params |
|---|---|---|
| FIRST | first subword unit | none |
| LAST | last subword unit | none |
| LAST2 | concatenation of the last two subword units | none |
| F+L | $wu_{\text{first}} + (1-w)u_{\text{last}}$ | $w$ |
| SUM | elementwise sum | none |
| MAX | elementwise max | none |
| AVG | elementwise average | none |
| ATTN | Attention over the subwords, weights generated by an MLP | MLP |
| LSTM | biLSTM reads all vectors, final hidden state | LSTM |

# Usual subword tokenisers

| BPE | Word Piece | Unigram Model |
| --- | --- | --- |
| co w o r k ing | c ow o r k i ng | coworking |
| co w orking | cowork ing | cowork ing |
| start: single characters<br>merge: most frequent<br>word boundary: end | start: single characters<br>merge: MI<br>word boundary: beginning | start: all hypotheses<br>split: max log-likelihood<br>word boundary: NA |

# Alternative BPE




**BPE**

co w o r k ing

co work ing

Gutierrez et al.

EACL 2021