Attending to Future Tokens for Bidirectional Sequence Generation: Supplementary Material

Carolin Lawrence and **Bhushan Kotnis** and **Mathias Niepert**
NEC Laboratories Europe
{carolin.lawrence, bhushan.kotnis, mathias.niepert}@neclab.eu

A Forward & Backward Attention

Figures 1 and 2 present the normalized forward attention $\bar{\alpha}^2$ and backward attention $\bar{\alpha}^3$ for the different attention heads over the sequence generation part for the SHARC and DAILY DIALOG dataset, respectively.

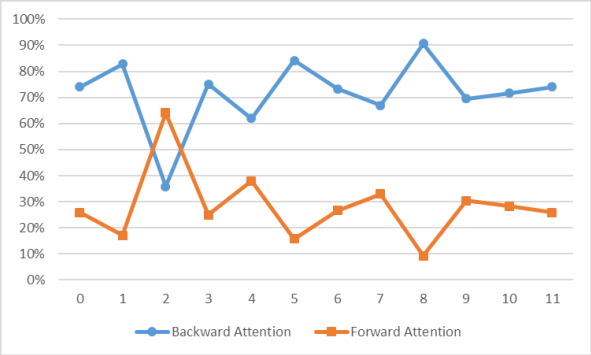

Figure 1: $\bar{\alpha}^2$ (Backward Attention) and $\bar{\alpha}^3$ (Forward Attention) across the 12 attention heads for the SHARC dataset.

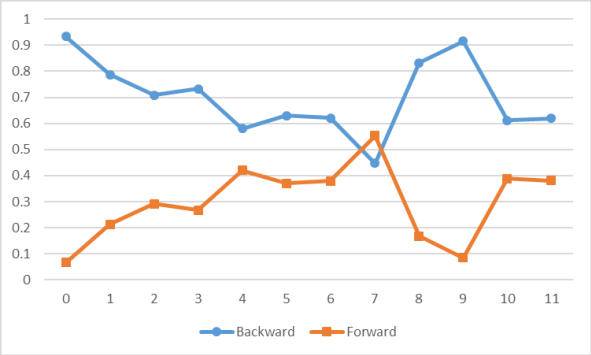

Figure 2: $\bar{\alpha}^2$ (Backward Attention) and $\bar{\alpha}^3$ (Forward Attention) across the 12 attention heads for the DAILY DIALOG dataset.

B Heat Maps

The heat maps (darker hues indicate higher attention) of Figures 3 and 4 show examples of where an attention head strongly looks into the future while generating from left to right. Each row shows the attention over the output sequence for this row's placeholder token at that point in time. Word in previous rows have been produced already, whereas words of later rows still hold placeholder tokens. Thus the upper triangle of the matrix shows the attention that is paid to future tokens. The red square marks the point of interest. Best viewed in color.

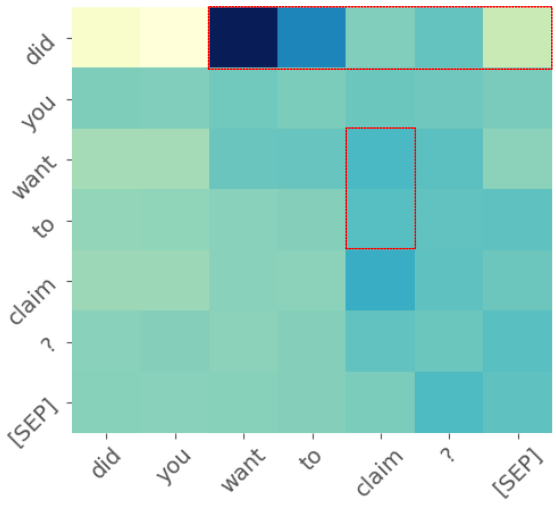

Figure 3

In Figure 3, we can see that when deciding on the first word, "*did*", high attention is paid to important future tokens. In particular, there is a strong focus on the 3rd placeholder, which is later revealed to be "*want*". Attention is also paid to the position that is later revealed to become a question mark. This indicates that model plans ahead and realizing a question will be formed, the word

"*did*" is chosen. Note that the similar sentence, "you want to claim." would not require the word "*did*" as it would not be a question.

Also in Figure 3, both the words "*want*" and "*to*" pay strong attention to the final word "*claim*" in the phrase "*want to claim*".

In Figure 4, when producing the word "*ambulance*" the attention is focused on the next placeholder token, which is in the next step revealed to be the word "*driver*" in the noun phrase "*ambulance driver*".

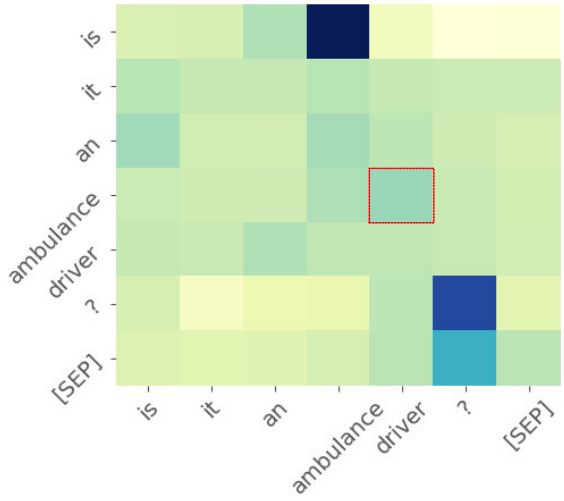

Figure 4