

A serial architecture

Try all candidate next words one at a time.

logit score for the candidate word

This allows the learned feature vector representation to be used for the candidate word.

hidden units that discover good or bad combinations of features

learned distributed encoding of word $t-2$

learned distributed encoding of word $t-1$

learned distributed encoding of **candidate**

↑ table look-up

↑ table look-up

↑ table look-up

index of word at $t-2$

index of word at $t-1$

index of candidate