A Supplemental Material

We also evaluate our work using Consecutive Wait (CW) as latency metrics, which measures the average lengths of consecutive wait segments. We also perform experiments on German↔English parallel corpora available from WMT15[3]. We use newstest-2013 as our dev set and newstest-2015 as our test set.

Fig. 8 show the translation quality on German↔English against AL of different decoding methods. Consistent to the results of Chinese↔English, our proposed speculative beam search gain large performance boost especially on test-time wait-k. Fig. 9 and Fig. 10 use CW as latency metrics. Since both the wait-k and test-time wait-k models use the same fixed policy, the CW latencies of the same k are identical.

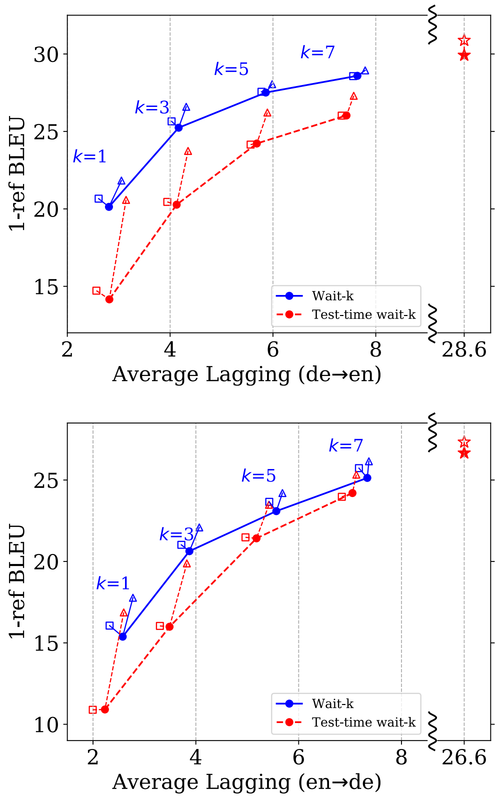

Figure 8: Translation quality against AL on English↔German simultaneous translation using wait-k model. □ □: conventional beam search only on target tail. △ △: speculative beam search. ★☆:full-sentence (greedy and beam-search).

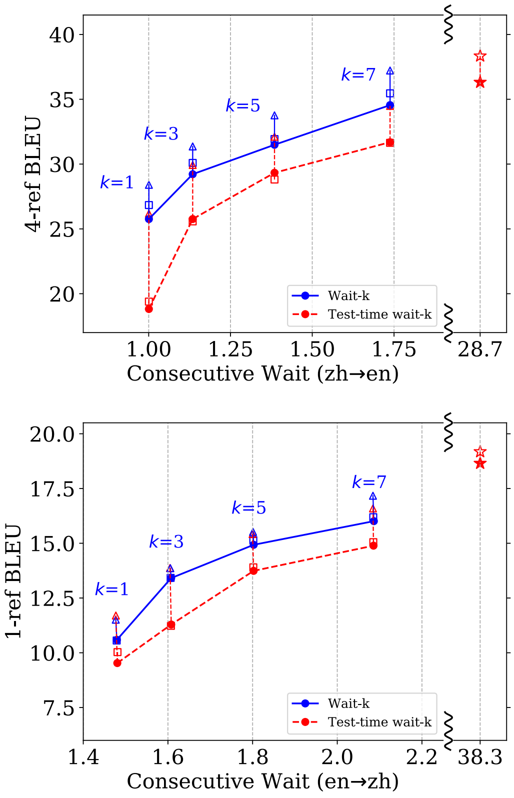

Figure 9: Translation quality against CW on Chinese↔English simultaneous translation using wait-k model.

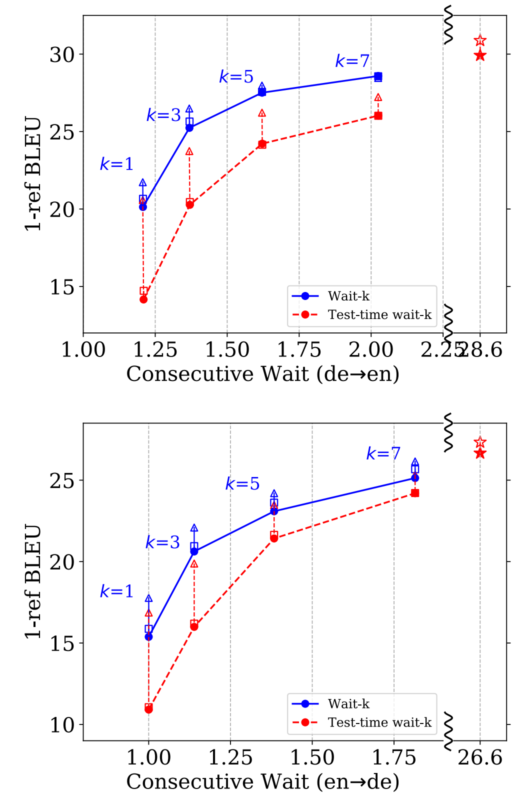

Figure 10: Translation quality against CW on English↔German simultaneous translation using wait-k model.

[3] http://www.statmt.org/wmt15/translation-task.html