Due to some problems with using HPC for jobs with large memory, we run training NN (continue baseline training, from **13** to **20** epochs) only one week ago. We found out that for this job 70 GB is not enough. 85 GB was OK.

Results were slightly better (we looked at perplexity and accuracy). Current best perplexity is **4.02**. What is more, perplexity converges, which means that there is no need to run more on this data and architecture.

With this model, we are now in progress of testing optimal value of beam_size parameter. We are searching through grid with values {1, 3, 5, …,19, 21} and storing BLEU scores.

We are planning either to try different architectures (1-2), tune best value of beam_size and then stack these models into ensemble, or train model on default architecture, but on subsets of training data (probably 4-5), but with this approach we need to know source (domain of texts) of dev and train data. We think that training on smaller data that have closer domain will improve results.