Natalia Rutecka

NLP - report of practical 5

e) The model with the greedy generation achieved the following results:

- Validation set: BLUE SCORE ≈ 0.241, corpus matches $\approx 80.61\%$, corpus success $\approx 56.12\%$
- Test set : BLUE SCORE ≈ 0.242, corpus matches $\approx 83.67\%$, corpus success $\approx 48.98\%$

The model with the beam search generation achieved similar results:

- Validation set: BLUE SCORE ≈ 0.240, corpus matches $\approx 80.61\%$, corpus success $\approx 57.14\%$
- Test set : BLUE SCORE ≈ 0.240, corpus matches $\approx 81.63\%$, corpus success $\approx 50.00\%$

f) The results of the greedy model:

- Validation set: BLUE SCORE ≈ 0.236, corpus matches $\approx 92.86\%$, corpus success $\approx 59.18\%$
- Test set : BLUE SCORE ≈ 0.214, corpus matches $\approx 95.92\%$, corpus success $\approx 53.06\%$

The results of the beam search model:

- Validation set: BLUE SCORE ≈ 0.233, corpus matches $\approx 93.88\%$, corpus success $\approx 67.35\%$
- Test set : BLUE SCORE ≈ 0.212, corpus matches $\approx 95.92\%$, corpus success $\approx 59.18\%$

The model with the softmax policy seems to perform better in terms of the inform rates and slightly better in terms of the success rates on both the validation and the test set, while having similar BLUE scores.

g) Using BLUE score shouldn't be the only way to evaluate a dialogue system, because it focuses solely on the similarity of the response to the reference ones, which doesn't always accurately represent the similarity in meaning. Moreover, some answers that are typically not at all similar to the reference ones, e.g. asking for clarification or replying that the model doesn't know something can be quite natural in the conversation, contrary to incoherent answers. The BLUE score also doesn't allow the model to learn to avoid biased answers.