Practical No. 5

Tomasz Siłkowski

ts407106@students.mimuw.edu.pl

3 maja 2023

2 Neural Implementation

e) Training and testing

After 15 epochs of training, following scores were achieved in testing:

Method	BLEU	Inform rate	Success rate
Greedy	0.236	95.92%	56.12%
Beam	0.239	96.94%	56.12%

f) Softmax Policy

After implementing softmax policy and then training it for 15 epochs, following scores were achieved in testing:

Method	BLEU	Inform rate	Success rate
Greedy	0.216	96.94%	45.92%
Beam	0.215	96.94%	44.90%

Although inform rate is slighty higher (and was trained faster), success rate is significantly lower when compared to default policy.

g) BLEU score as sole evaluation metric for dialogues

When BLEU metric is used in translation evaluation, it suffers from lack of checking for meaning and verifying the structure. In dialogue system, these problems are made worse, as we put more emphasis on integrity and comprehensiveness, especially when were trying to pass off as a human. If BLEU is used as a sole evaluation metric, results are more likely to have all the right components, but appear intensly unhuman, ruining the flow of conversation.