Practical No. 5

Jan Dziuba

May 11, 2023

1 Part 2

e)

Loading parameters of iter 10
————————————————————GREEDY
Current VALID LOSS: 0
Valid BLUES SCORE 0.2396838019
Valid Corpus Matches : 91.84%
Valid Corpus Success : 60.20%
Valid Total number of dialogues: 98
Current TEST LOSS: 0.0
Corpus BLUES SCORE 0.2428193872
Corpus Matches : 93.88%
Corpus Success : 57.14%
Total number of dialogues: 98
————————————————————BEAM
Current VALID LOSS: 0
Valid BLUES SCORE 0.2431330542
Valid Corpus Matches : 91.84%
Valid Corpus Success : 64.29%
Valid Total number of dialogues: 98
Current TEST LOSS: 0.0
Corpus BLUES SCORE 0.2476875225
Corpus Matches : 93.88%
Corpus Success : 60.20%
Total number of dialogues: 98
TIME: 3.9095523357391357

f)

Loading parameters of iter 10
————————————————————GREEDY
Current VALID LOSS: 0

Valid BLUES SCORE 0.2402353340
Valid Corpus Matches : 86.73%
Valid Corpus Success : 63.27%
Valid Total number of dialogues: 98
Current TEST LOSS: 0.0
Corpus BLUES SCORE 0.2325164968
Corpus Matches : 81.63%
Corpus Success : 48.98%
Total number of dialogues: 98
————————————————————BEAM
Current VALID LOSS: 0
Valid BLUES SCORE 0.2376763596
Valid Corpus Matches : 86.73%
Valid Corpus Success : 63.27%
Valid Total number of dialogues: 98
Current TEST LOSS: 0.0
Corpus BLUES SCORE 0.2365194855
Corpus Matches : 81.63%
Corpus Success : 50.00%
Total number of dialogues: 98
TIME: 4.103184700012207
After train and test rerun
Loading parameters of iter 10
————————————————————GREEDY
Current VALID LOSS: 0
Valid BLUES SCORE 0.2502195127
Valid Corpus Matches : 91.84%
Valid Corpus Success : 57.14%
Valid Total number of dialogues: 98
Current TEST LOSS: 0.0
Corpus BLUES SCORE 0.2394296606
Corpus Matches : 86.73%
Corpus Success : 52.04%
Total number of dialogues: 98
————————————————————BEAM
Current VALID LOSS: 0
Valid BLUES SCORE 0.2507824854
Valid Corpus Matches : 91.84%
Valid Corpus Success : 57.14%
Valid Total number of dialogues: 98
Current TEST LOSS: 0.0
Corpus BLUES SCORE 0.2440719282
Corpus Matches : 87.76%
Corpus Success : 53.06%
Total number of dialogues: 98

TIME: 4.0200865268707275
Surprisingly results are worse with softmax policy.
After retraining results improve, but are still worse than default policy.

g)

BLEU score is a quick and simple way of evaluating dialogue systems. It shouldn't be the only metric for evaluating the model as it doesn't take into account synonymous words.