Practical No. 5

Emilia Wiśnios
ew407219@students.mimuw.edu.pl

May 17, 2022

Part 2

e)

Obtained scores from the Vanilla model (epoch 13):

- Greedy
  - Valid BLUES SCORE : 0.2418071330
  - Valid Corpus Matches : 92.86%
  - Valid Corpus Success : 54.08%
  - Corpus BLUES SCORE : 0.2322883255
  - Corpus Matches : 95.92%
  - Corpus Success : 39.80%
- BEAM
  - Valid BLUES SCORE : 0.2390998134
  - Valid Corpus Matches : 93.88%
  - Valid Corpus Success : 57.14%
  - Corpus BLUES SCORE : 0.2232929454
  - Corpus Matches : 98.98%
  - Corpus Success : 41.84%

f)

Obtained scores from the Softmax model (epoch 13):

- Greedy
  - Valid BLUES SCORE : 0.2426519285

- Valid Corpus Matches : 95.92%
- Valid Corpus Success : 72.45%
- Corpus BLUES SCORE : 0.2336958274
- Corpus Matches : 84.69%
- Corpus Success : 54.08%

- BEAM
  - Valid BLUES SCORE : 0.2256414222
  - Valid Corpus Matches : 94.90%
  - Valid Corpus Success : 70.41%
  - Corpus BLUES SCORE : 0.2232929454
  - Corpus Matches : 85.71%
  - Corpus Success : 53.06%

Corpus success is higher in the Softmax approach.

Part 3

a)

1. - Ground truth

     I am sorry, I could not find any Cuban restaurants for you in the north part of town. Is there a different type of food you would like me to search for instead?

   - Model's answer

     what kind of food are you looking for ?

     Model omitted the whole part informing the customer that there are no Cuban restaurants in given area. To fix this problem we should increase training set size and train the model for more epochs. We should also consider adding pre-trained model.

2. - Ground truth

     Saigon City is located in the north part of town and serves Asian Oriental food. Would you like their address?

   - Model's answer

     there are [value_ count] restaurant -s in the [value_ area] part of town . would you like ?

     The sentence is not finished. To fix this problem we should increase training set size and train the model for more epochs. We should also consider adding pre-trained model.

3.
   - Ground truth
     do you want particular food type?
   - Model's answer
     [restaurant_name] is a [value_food] restaurant in the [value_area] part of town
     The model is proposing restaurants without having the full information about customer's needs. To fix this problem we should increase training set size and train the model for more epochs. We should also consider adding pre-trained model.

b)

Automatic metrics as BLEU are convenient for fast, efficient and reproducible research. Nevertheless those metric have been shown to only "correlate very weakly with human judgement". Due to open-ended nature of conversations, there are may possible responses in a given dialogue, and there is typically only one ground truth label available. Thus I don't think that one should evaluate the model only relying on this metric. [4]

c)

Our model has more than 100k trainable parameters. We used 676 dialogues and 2744 turns (question and answers) to train the model. This ratio of parameters may be problematic – the model would probably overfit to the structure of dialogues. We can see that in section a) – the model learned some fixed phrases. Responses given by this model are dull and repetitive. According to Google Developers Blog *As a rough rule of thumb, your model should train on at least an order of magnitude more examples than trainable parameters.*

d)

Data collection for task-oriented conversations are very costly and time consuming. Unless external resources happen to already be available (which is not the case for most domains), in-domain data collection requires having a deployed system capable of sustaining a dialog with a user. This leads to bootstrapping problem: without data to train the initial system, the developers either use datasets like Wizard-of-Oz or try to manually develop grammars and language models. Collecting dialog data with an early version of a deployed system has shortcomings: data collection quality may suffer from the inadequacies of the system itself, and users may bias their language to adjust for the deficiencies of the system in the pursuit of having a successful dialog. Finally, the whole process must be repeated all over again for each new domain or system, or when the new functionality is added. [3]

e)

The current approaches to build task-oriented dialog systems still require a substantial amount of annotations and therefore are labor-intensive. On the other hand, large-scale

pre-trained language models such as BERT and GPT have achieved great success on various NLP tasks, which proves the effectiveness of pre-training. [2] Apart from that, using pre-trained models improve the quality of generated answers. Generative models like GPT are now able to create longer and more coherent sequence outputs, which is a remedy for problem of dull and repetitive response generation. One other thing are out-of-vocabulary word – long standing challange in dialogue modeling, which can be solved with pre-trained models. [1] A variety of datasets could be utilized – articles from the internet, wikipedia, (not to old) books.

References

[1] Paweł Budzianowski and Ivan Vulić. Hello, It's GPT-2 – How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems, 2019.

[2] Jing Gu, Qingyang Wu, Chongruo Wu, Weiyan Shi, and Zhou Yu. A Tailored Pre-Training Model for Task-Oriented Dialog Generation, 2020.

[3] Walter Lasecki, Ece Kamar, and Dan Bohus. conversations in the crowd: Collecting data for task-oriented dialog learning.

[4] Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. Human Evaluation of Conversations is an Open Problem: comparing the sensitivity of various methods for evaluating dialogue agents. *CoRR*, abs/2201.04723, 2022.