Week 3 - Dataset exploration

20 October 2021 15:18

The task in this lab is to explore dialogue datasets and find out more about them. Your job will thus be to write a script that computes some basic statistics about datasets, and then try to interpret the script's results.

1. We provide the dialog dataset in .txt format (in tab "files", named "dialog-babi-task5-full-dialogs-trn.txt"). This file contains data from testing end-to-end dialog systems in the restaurant domain, described in the following paper:

   Antoine Bordes, Y-Lan Boureau, Jason Weston, Learning End-to-End Goal-Oriented Dialog, arxiv:1605.07683.

   The file format for each task is as follows:

   ID user_utterance [tab] bot_utterance
   ...

   The IDs for a given dialog start at 1 and increase. When the IDs in a file reset back to 1 you can consider the following sentences as a new dialog. When the bot speaks two times in a row, we used the special token “<SILENCE>” to fill in for the missing user utterance. See more details in the README included with the dataset. Bot utterances can be sentences or API calls (sentences starting with the special token “api_call”). Here is an example of dialog:

1 hi	hello what can i help you with today

2 can you make a restaurant reservation with italian cuisine for six people in a cheap price range	i'm on it

3 <SILENCE>	where should it be

4 rome please	ok let me look into some options for you

5 <SILENCE>	api_call italian rome six cheap

   Green stands for dialog turns, yellow for user utterances and blue for bot utterances. User and bot utterances are separated by a **tab,** while all other words by space.

2. Write a script that will **read all turns in the data and separate the user and system utterances** in the training set.
   - Make the script **ignore any search results lines** in the data (they **don't** contain a tab character). For example:

     1 good morning hello what can i help you with today

     2 i'd like to book a table with italian food i'm on it

     3 <SILENCE> where should it be

4 in paris how many people would be in your party

5 for six people please which price range are looking for

6 in a cheap price range please ok let me look into some options for you

7 <SILENCE> api_call italian paris six cheap

8 actually i would prefer for two people sure is there anything else to update

9 instead could it be in madrid sure is there anything else to update

10 instead could it be with spanish food sure is there anything else to update

11 no ok let me look into some options for you

12 <SILENCE> api_call spanish madrid two cheap

13 resto_madrid_cheap_spanish_1stars R_phone resto_madrid_cheap_spanish_1stars_phone

14 resto_madrid_cheap_spanish_1stars R_cuisine spanish

15 resto_madrid_cheap_spanish_1stars R_address resto_madrid_cheap_spanish_1stars_address

16 resto_madrid_cheap_spanish_1stars R_location madrid

17 resto_madrid_cheap_spanish_1stars R_number two

18 resto_madrid_cheap_spanish_1stars R_price cheap

19 resto_madrid_cheap_spanish_1stars R_rating 1

20 resto_madrid_cheap_spanish_6stars R_phone resto_madrid_cheap_spanish_6stars_phone

21 resto_madrid_cheap_spanish_6stars R_cuisine spanish

22 resto_madrid_cheap_spanish_6stars R_address resto_madrid_cheap_spanish_6stars_address

23 resto_madrid_cheap_spanish_6stars R_location madrid

24 resto_madrid_cheap_spanish_6stars R_number two

25 resto_madrid_cheap_spanish_6stars R_price cheap

26 resto_madrid_cheap_spanish_6stars R_rating 6

27 resto_madrid_cheap_spanish_8stars R_phone resto_madrid_cheap_spanish_8stars_phone

28 resto_madrid_cheap_spanish_8stars R_cuisine spanish

29 resto_madrid_cheap_spanish_8stars R_address resto_madrid_cheap_spanish_8stars_address

30 resto_madrid_cheap_spanish_8stars R_location madrid

31 resto_madrid_cheap_spanish_8stars R_number two

32 resto_madrid_cheap_spanish_8stars R_price cheap

33 resto_madrid_cheap_spanish_8stars R_rating 8

34 <SILENCE> what do you think of this option: resto_madrid_cheap_spanish_8stars

35 no this does not work for me sure let me find an other option for you

36 <SILENCE> what do you think of this option: resto_madrid_cheap_spanish_6stars

37 do you have something else sure let me find an other option for you

38 <SILENCE> what do you think of this option: resto_madrid_cheap_spanish_1stars

39 it's perfect great let me do the reservation

40 may i have the phone number of the restaurant here it is resto_madrid_cheap_spanish_1stars_phone

41 thanks is there anything i can help you with

42 no thank you you're welcome

You should ignore the grey lines (TIP: search if the line contains a tab. If not, then ignore.)

- If the script finds a turn where the user is silent (the user turn contains only `<SILENCE>`), it should concatenate the system response from this turn to the previous turn. Note that this may happen on multiple consecutive turns, and the script should join all of these together into one system response. For example, for lines 35 and 36, user utterance is: *no this does not work for me* and system response is: *sure let me find an other option for you what do you think of this option: resto_madrid_cheap_spanish_6stars* (concatenated system responses)

- If `<SILENCE>` is the first word in the dialogue, just delete it.

- Don't worry too much about tokenization (word segmentation) -- tokenizing on whitespace is OK.

1. Implement a routine that will **compute the following statistics** for system and user turns (separately):
    - data length (total number of dialogues, turns, words)
    - mean and standard deviations for individual dialogue lengths (number of turns in a dialogue, number of words in a turn)
    - vocabulary size

    Commit this file as `hw2/stats.py`.