DL4NLP Shared Task 2021

June 1, 2021

Contents

1 Task

The topic of the DL4NLP 2021 shared task deals with conversational agents. More specific, your task is to design a model that can rank a list of possible clarifying questions given an initial user query. You will be given a training set as well as a small example test set and a file that contains all possible clarifying questions. This task can be tackled by designing a classification model or a generative model. **Your task is to design a model that ranks a list of possible clarifying questions and return the best suiting ones.**

2 Dataset and Evaluation

2.1 Dataset

The dataset contains a training.tsv file which contains user queries and a list of possible clarifying questions for the query. Note that a single query has multiple possible clarifying questions (one query, multiple rows in the training set) and in this task we treat them all as a "correct" clarifying question for the given query even though you may find that some clarifying questions fit better than others. Since the training.tsv only contains positive training samples ("Correct" clarifying question for the query) you will have to create the negative training samples yourself by combining queries with random clarifying questions. You will also have to split the training.tsv and use a subset of the training data as validation/dev data to validate your model. The question_bank.tsv file contains a list of all possible clarifying questions.
The test.tsv file contains a few example test questions that you can use to test your model on unseen data. There won't be a big set of test queries because it takes a bit of time to evaluate one query against every clarifying question.

2.2 Metric

Your models performance will be evaluated in top-k accuracy. For each query in the test data you need to return the top-50 clarifying questions ranked by your model. If any of the top 50 questions contains **any** of the correct ground truth clarifying question, that query will be considered as correct. For each query in the test set you should return **one** list of top-50 clarifying questions. The whole test set will be given in the development phase of the shared task. This test set won't contain the corresponding ground truth

clarifying questions. You will also be given an example submission at the start of that phase.
The test queries only have one or a few possible clarifying questions to make things easier.

3 Technical Details

3.1 Frameworks and models

For this shared task you may use any framework and any (pretrained)model that you like. You can tackle this shared task in any way you want. For example, you could use a single BERT model, use the query and a clarifying question as input separated by a separator token and use the NSP/CLS token as a classification token whether this combination of query and clarifying question is fitting. You could even create a generative model and rank the clarifying questions by likelihood.
The only restriction is that it has to be a deep learning model and your model needs to be able to classify each query and clarifying question combination on its own (No softmax output over all the questions in the question bank).

3.2 Nice to have resources

Doing the task without a GPU isn't really possible due to the amount of data. Therefore you should use your GPU or, if you don't have access to a GPU, you can use GoogleColab [1] to have free access to a GPU. You may find the huggingface transformer repo [2] useuful as well as the UKPLabs sentence transformer repo [3].

4 Organizational Matters

4.1 Groups

You will have a separate group selection for the shared task. The Groups are limited to 4 people with the intention that two homework groups can work together. Please only select a group if you want to do the shared task.

4.2 Phases

The shared task is split into three phases:

1. In the warmup phase you can develop the first model and get familiar with the task. You will also be able to select your group in this phase. This phase will last from **1.6.2021 to 8.6.2021**
2. In the development and test phase you will be able to submit predictions for the test set that will be released when this phase starts. You will submit your predictions on CodaLab once it is set up. This phase lasts from **8.6.2021 - 13.7.2021**.
3. The report phase lasts from **1.6.2021 - 13.7.2021**. You will have to submit a small report about your approach until then.

4.3 Cheating

Of course, we are totally aware that you could simply look up the best suiting questions in the questions bank yourself to boost your performance. If your models description and performance doesn't seem to match up we may ask you to submit your code to verify the results.

4.4 Grading

For getting the points for the shared task you will have to have 35% of the test set correct, meaning for 35% of the test queries you will have to have one of the correct ground truth clarifying questions in the top-50 predictions of your model. The test data is designed to have quite easy hints for the model to find the correct clarifying question to make the task not too complicated. There are 50 possible points to achieve in the shared task. You will get 25 points for reaching the 35% on the test set and 25 points for a well written report. Therefore, even if you don't reach the 35% you may want to write a report about your approach.

[1] https://colab.research.google.com/notebooks/intro.ipynb?utm_source=scs-index
[2] https://github.com/huggingface/transformers
[3] https://github.com/UKPLab/sentence-transformers

5 Written Report

To document your approach you will have to submit a written report with 2-4 pages. It must be submitted in the ACL 2020 [4] [5] template in a pdf format. The paper must include the following points:

- Introduction
- Description of your approach
- How you split your training.tsv into a train and dev set
- How you generated negative samples
- Results and discussion of those

[4] https://www.overleaf.com/latex/templates/acl-2020-proceedings-template/zsrkcwjptpc

[5] http://acl2020.org/downloads/acl2020-templates.zip