Dialogue and Narrative 2021: Coursework Assignment

Your goal is to implement and evaluate a dialogue system that can help users with an information-seeking task. For this coursework, you can choose to focus on one of two sub-tasks from the DialDoc21 shared task:

Subtask 1: Identify the knowledge in an associated document that the agent needs to respond to the user. The input is the dialogue history and current user turn (all the dialogue uttered so far), and the associated document. The output is a text span.

Subtask 2: Generate a response using the knowledge extracted from the document. The input is the dialogue history, current user turn and the extracted text span.

These two tasks will use the techniques and skills we have been testing in our Wednesday lab sessions. For this assignment, you don't need to worry about 'winning' the competition. Instead, focus on developing your understanding of designing, building, and evaluating an NLP system. Rather than a novel method, we are looking for a scientific approach to designing and evaluating the system and a technically sound implementation, based on what we have learned during the unit.

Submission

The work will be assessed by a report in the style of a short research paper,

Due on: 12th January 2022, at 1pm.

A typical report will include the motivation for your approach, a brief technical background, a description of your method, your experimental setup and results, and some conclusions that share any insights you gained about how the system works.

Format: 4–8 page report **(excluding references)**. We suggest you follow the SemEval formatting guidelines by using these style files. You may alternatively use the guidelines here but note that the sections described in these guidelines are just suggestions.

Teamwork: you can work either alone or in teams of two people. If you're in a team, you only need to submit one report, but it should reflect double the amount of work. Both people must be involved in the implementation and testing as well as writing up.

Support

The Wednesday lab sessions will run from weeks 8-12 as coursework help sessions. Please post any questions to the Blackboard discussion board, our unit Team, or email Edwin.simpson@bristol.ac.uk.

Some Hints

- Build a working solution to the task:
    - Start with the simplest thing that (kind of) works -- a ‘baseline’ system -- then try to improve it.
    - Note that methods that require training a deep neural network require a lot of GPU time to train (from several hours to several days). You may wish to use Google Colab for this.
    - Make use of existing libraries, such as HuggingFace Transformers, AllenNLP, Scikit-learn, Spacy, NLTK, Gensim or Flair. There are many tutorials and examples available online that show how to solve related tasks using these libraries.
- In your report, you can provide various insights about your system:
    - Which parts of your system are most time-consuming to develop or to run?
    - If the system involves several processing or preprocessing steps, what kind of errors arise in step?
    - Give some examples of where your method gets the wrong answer and explain what the likely cause is — provide ‘error analysis’.
    - What mistakes does the system make that a human would not?
- Follow a scientific approach with clear hypotheses:
    - You should have a clear hypothesis about how well your approach will work and why, and your evaluation and error analysis should test this hypothesis.
    - Start with a relatively simple ‘baseline’ method, then consider how you could improve it, e.g., by improving the preprocessing steps, the embeddings, or the machine learning method.
- Non-deep learning approaches may be interesting to compare with results from previous work. Often, shallow methods can provide surprisingly strong performance and help us to understand the complexities of a particular task and dataset.

Marking Criteria

70% or above:

- The motivation and technical background are concise but show a strong understanding of the chosen techniques.
- Design decisions are well justified, and the system uses a range of different NLP techniques, which are implemented and applied correctly.
- The implementation provides a functioning system for your chosen subtask and good performance relative to baselines.
- The experimental evaluation is sound and draws clear conclusions about the strengths and weaknesses of the tested method(s).
- The evaluation includes a detailed analysis, such as examples of specific types of errors or ablation tests with accompanying discussions.
- The report shows original thinking and synthesis of new ideas beyond those covered by the required reading.
- Very good presentation of work, with a clear structure, descriptions and excellent use of plots or tables to communicate the results.

60-70%:

- The motivation and technical background show a good understanding of the chosen techniques.
- Design decisions are justified, and the system uses several NLP techniques, which are largely implemented and applied correctly.
- The implementation provides a functioning system for your chosen subtask.
- The experimental evaluation is mostly sound and the student was able to draw some conclusions about the strengths and weaknesses of the tested method(s).
- The evaluation includes some additional analysis beyond performance metrics, such as discussion of examples of generated dialogue.
- The report shows a very good grasp of the topics studied in the reading group.
- Good presentation of work, with a fairly clear structure, descriptions and appropriate use of plots or tables to communicate the results.

50-60%

- The motivation and technical background show some understanding of the chosen techniques.
- Design decisions are only partially justified, and the system uses a very limited set of NLP techniques, with a few minor mistakes in the implementation.
- The implementation provides a partially functioning system for your chosen subtask.
- The experimental evaluation has some flaws but is still able to draw some conclusions about the strengths and weaknesses of the tested method(s).
- There is little further analysis beyond performance metrics and limited critical judgement of the methods.
- The report shows a good grasp of some topics studied in the reading group.
- Fairly good presentation of work, with some flaws in the structure of the report, omissions in the descriptions or shortcomings in the presentation of results.

Below 50% (fail)

- The motivation and technical background show very limited understanding of NLP techniques.
- Design decisions are not justified, and the system uses a very limited set of NLP techniques, with notable mistakes in the implementation.
- Major steps of the implementation are incomplete, and the system cannot be evaluated on the chosen subtask.
- The experimental evaluation has notable flaws that limit the conclusions we can draw.
- There is no further analysis or critique beyond performance metrics.
- The report provides limited evidence of understanding of the topics studied in the reading group.
- There are presentation issues in the report, such as confusing structure, missing important information, or a lack of results.