Conversational Interfaces Coursework 2023

Q2: NLU Evaluation (5 marks)

In this question you will run a set of test user utterances on your created agent to evaluate the NLU effectiveness.

A set of 13 sample test dialogues are provided (in file **test-data-updated.zip** on Moodle) for you to test your NLU **manually**, as described below. You should use your selected tool's test process to evaluate the NLU effectiveness on the user utterances from these dialogues. The focus of this evaluation will be on the performance of the **GetInfo** intent.

Each dialogue is presented in the same format as the original data. For example, here is the start of the dialogue in the file **dlg-4a1341e7-cf17-4dd3-9d9c-620ae9f10c81.txt:**

```
ASSISTANT: Hello.

USER: Hi. Today, I need your help getting information about Aston Villa in EPL. I'm
wondering, what is their record right now?
- Intent: GetInfo
- Slots: {team: 'Aston Villa', info: 'winLossRecord'}

ASSISTANT: Aston Villa's record is currently in 16th place in the Premier League.

USER: Okay. Now I'm wondering, Aston Villa in their last game, what was the score?
- Intent: GetInfo
- Slots: {team: 'Aston Villa', info: 'lastScore'}

ASSISTANT: It was their first game.
```

You should feed **each user turn** in the test dialogues to your system's NLU, using the appropriate test harness (e.g., Alexa's "Test" interface), and record the intent and slots that are predicted for each turn. The focus is on the NLU output for each individual utterance; you do not need to worry about whether your system responses are similar to those in the examples, or whether the dialogues make sense as a whole.

You should then use the predicted slots and intents to compute the following measures:

- Intent Precision, Recall, and F1 score. For this measure, you should consider any intents that are not **GetInfo** to be **Other**.
- Average Slot Precision, Recall, and F1 score – this should be computed for all slots over all eligible turns. Eligible turns are any that are labelled as **GetInfo** in the ground truth and/or the NLU prediction

You are not required to update your system in any way in response to the results of this evaluation. You will be marked on whether you carried out the evaluation correctly and report the numbers appropriately, not on the actual results.

What to submit

Report the results on all of the above metrics in a table. Discuss the overall effectiveness of your NLU, and discuss any reasons that might explain your system performance, giving examples from the test dialogues. Include in an appendix the NLU labels for each user turn in the test dialogues.