Spoken Language Understanding in Task-oriented Dialog Systems

Overview

In task-oriented dialog systems, understanding of users' queries (expressed in natural language) is a process of parsing users' queries and converting them into some structure that machine can handle. The understanding usually consists of two parts, namely intent identification and slot filling. The textual strings, fed into a dialog system as input, are mostly the transcripts translated from spoken language by ASR (Automatic Speech Recognition) and thus subject to recognition errors.

The dataset adopted by this task is a sample of the real query log from a commercial task-oriented dialog system. The data is all in Chinese. The evaluation includes three domains, namely music, navigation and phone call. Within the dataset, an additional domain label 'OTHERS' is used to annotate the data not covered by the three domains. To simplify the task, we keep only the intents and the slots of high-frequency while ignoring others although they appear in the original data. The entire data can be seen as a stream of user queries ordered by time stamp. The stream is further split into a series of segments according to the gaps of time stamps between queries and each segment is denoted as a 'session'. The contexts within a session are taken into consideration when a query within the session was annotated. Below are two example sessions with annotations.

```
1	打电话	phone_call.make_a_phone_call	打电话
1	我想听美观	music.play	我想听<song>美观</song>
1	我想听什话	music.play	我想听<song>什话||神话</song>
1	神话	music.play	<song>神话</song>
2	播放调频广播	OTHERS	播放调频广播
2	给我唱一首一晃就老了	music.play	给我唱一首<song>一晃就老了</song>
```

Format:

- ✧ Each line consists of four fields separated by '\t'. They are session ID, user query, intent, and slot annotation.
- ✧ The sessions include only user queries, not system responses.
- ✧ The queries in sessions are ordered by timestamp.
- ✧ The corrected values of the slot values are included in annotations as well if the slot values contains ASR errors. For example, for the slot annotation "我想听<song>什话||神话</song>", the string "神话" is correction of "什话".

To help participating systems correct ASR errors, this task also provides a dictionary of values for each type of slot. Note that dictionaries are pruned such that they include all the values occurring in the evaluation dataset, but do not necessarily include all the values in real world.

Task Description

The task consists of four sub-tasks. The participating teams can optionally output results for some sub-tasks, not necessarily for all of them.

Sub-task 1: Intent Identification – Close

For this sub-task, the participating systems are required to

- ✧ use only the training dataset provided by the task for the model training/tuning of intent identification, and
- ✧ output the results (in the evaluation stage) based only on the provided test set, not on any other dataset or resources.

The evaluation metric is $F1_{macro}$, calculated as the following equations,

$$P_{macro} = \frac{1}{N}\sum_{i=1}^{N} \frac{\#\text{ of queries correctly predicted as intent } c_i}{\#\text{ of queries predicted as intent } c_i},$$

$$R_{macro} = \frac{1}{N}\sum_{i=1}^{N} \frac{\#\text{ of queries correctly predicted as intent } c_i}{\#\text{ of queries labelled as intent } c_i},$$

$$F1_{macro} = \frac{2}{\frac{1}{P_{macro}} + \frac{1}{R_{macro}}}$$

Sub-task 2: Intent Identification – Open

For this sub-task, the participating systems

- ✧ can use any datasets and resources (in addition to the provided training dataset) for model training/tuning of intent identification, and
- ✧ are required to output the results (in the evaluation stage) based only on the provided test set, not on any other dataset or resources.

The evaluation metric is $F1_{macro}$, same as sub-task 1.

Sub-task 3: Intent Identification and Slot Filling – Close

For this sub-task, the participating systems are required to

- ✧ use only the dataset and the slot dictionaries provided by the task for the model training/tuning of intent identification and slot filling, and

- ✧ output the results (in the evaluation stage) based only on the provided test set and the provided slot dictionaries, not on any other dataset or resources.

The evaluation metric is as given by the following equation.

$$P = \frac{\#\text{ of queries correctly parsed}}{\#\text{ of queries}}$$

where "# of queries" is the number of queries in the test set (including the queries with intent annotated as 'OTHERS'). "# of queries correctly parsed" denotes the number of queries for which the predicted intent and the predicted slot values (including the corrected values if correction is needed) are both exactly same as the annotations.

Sub-task 4: Intent Identification and Slot Filling - Open

For this sub-task, the participating systems

- ✧ can use any datasets (in addition to the training dataset and the slot dictionaries provided by the task) for the model training/tuning of intent identification and slot filling, and
- ✧ output the results (in the evaluation stage) based only on the provided test set and the provided slot dictionaries, not on any other dataset or resources.

The evaluation metric is P value, same as sub-task 3.