Training Data of DialEval-1 Task

Recently, many reserachers are trying to build automatic helpdesk systems. However, there are very few methods to evaluate such systems. In **DialEval-1**, we aim to explore methods to evaluate task-oriented, multi-round, textual dialogue systems automatically. This dataset have the following features:

- Chinese customer-helpdesk dialogues carwled from Weibo.
- English dialgoues: manually translated from a subset of the Chinese dialgoues.
- Nugget type annotatoins for each turn: indicate whether the current turn is useful to accomplish the task.
- Quality annotation for each dialogue.
    - task accomplishment
    - customer satisfaction
    - dialogue effectiveness

In DialEval-1, we consider annotations ground truth, and participants are required to predict nugget type for each turn (Nugget Detection, or ND) and dialogue quality for each dialogue (Dialogue Quality, or DQ).

Links

- Homepage of DialEval-1 Task
- Introduction of the training dataset (current page)
- NTCIR-15

Registration

TO and register and obtain the dataset ,please send an email to dialeval1org@list.waseda.jp with the following information so that we can send you the training data.

- Team Name (e.g. Waseda)
- Principal investigator's name, affilication, email address
- Names, affiliations, email addresses of other team members
- Subtasks that you plan to participate: Chinese, English, or BOTH

Later, NII will require you to register to NTCIR tasks through their website, but please contact us by email first

Leaderboard

http://deepimagining.com/leaderboard

Overview of Dataset

Training Data

The Chinese training dataset contains 4,090 (3,700 for training + 390 for dev) customer-helpdesk dialgoues which are crawled from Weibo. All of these dialogues are annotated by 19 annotators.

The English dataset contains 2, 251 dialogues for training + 390 for dev. They are manually translated from a subset of the Chinese dataset. The English dataset shares the same annotations with the Chinese dataset.

- Training set
  - train_cn.json (3,700 Chinese dialogues)
  - train_en.json (2, 251 English dialogues)
- Dev set
  - dev_cn.json (390 dialogues)
  - dev_en.json (390 dialogues)

Test Data

Test set

- test_cn.json (300 dialogues)
- test_en.json (300 dialogues)

Annotation

We hired 19 Chinese students to annotate the training/dev dataset in 2018. In 2019, the test dataset of DialEval-1 were annotated by another group of annotators. Thus, there may be a gap between the training data and test data, as the dialogue annotation is quite subjective.

Format of the JSON file

Each file is in JSON format with UTF-8 encoding.

Following are the top-level fields:

- **id**
- **turns**: array of turns from the customer and the helpdesk (see details below)
- **annotations**: a list of annotations provided by 19 annotators. Each annotation consists of two fields: **nugget** and **quality**

Each element of the turns field contains the following fields:

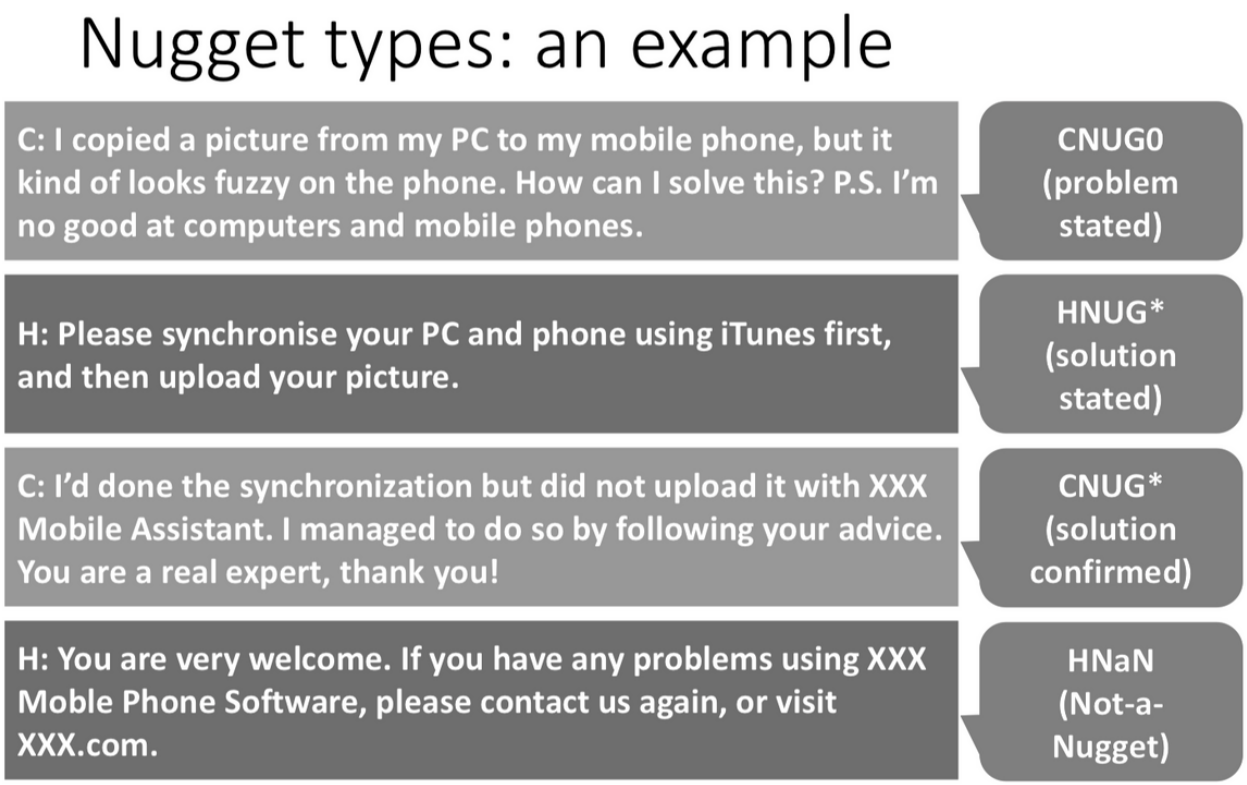
- **sender**: the speaker of this turn (either customer or helpdesk)
- **utterances**: the utterances (may be multiple) they sent in this turn. Note that some utterances are empty strings since we didn't crawl emoji and photos.

Each element of **annotations** contains the following fields:

- **nugget**: The list of nugget types for each turn (see details below).
- **quality**: A dictonary consists of the subjetive dialogue quality scores: A-score, S-score, and E-score (see details below).

Nugget Types

- CNUG0: Customer trigger (problem stated)
- CNUG*: Customer goal (solution confirmed)
- HNUG*: Helpdesk goal (solution stated)
- CNUG: Customer regular nugget
- HNUG: Helpdesk regular nugget
- CNaN: Customer Not-a-Nugget
- HNaN: Helpdesk Not-a-Nugget



Dialogue Quality

- A-score: Task **A**ccomplishment (Has the problem been solved? To what extent?)
- S-score: Customer **S**atisfaction of the dialogue (not of the product/service or the company)
- E-score: Dialogue **E**ffectiveness (Do the utterers interact effectively to solve the problem efficiently?)

Scale: [2, 1, 0, -1, -2]

Evaluation

Metrics

During the data annotaiton, we noticed that annotators' assessment on dialgoues are highly subjective and are hard to consolidate them into one gold label. Thus, we proposed to preserve the diverse views in the annotations “as is” and leverage them at the step of evaluation measure calculation.

Instead of juding whether the estimated label is equal to the gold label, we compare the difference between the estiamted distributions and the gold distributions calculaed by 19 anntators' annotations). Specifically, we employ these metrics for quality sub-task and nugget sub-task:

- Quality:
  - **NMD**: Normalised Match Distance.
  - **RSNOD**: Root Symmetric Normalised Order-aware Divergence
- Nugget:
  - **RNSS**: Root Normalised Sum of Squared errors
  - **JSD**: Jensen-Shannon divergence

For the details about the metrics, please vistit:

- NTCIR-15 Dialogue Evaluation Task Definition
- Comparing Two Binned Probability Distributions for Information Access Evaluation.

Test and Submission

`example_submission_format` contains json files which are expected submission format.

Each submission json is for a specific language and subtask, e.g. quality_english.json

We provide a online evaluation tool for your to evaluate your model on test data before submitting your final runs:

http://deepimagining.com/upload

To avoid over-fitting on test-data, only 50% of the test data are used by the online evaluation tool.

We will evaluate your final submission with 100% test data.

Please send your final submission to dialeval1org@list.waseda.jp by Jul 31

Schedule

- Jul 2019 Test data crawling [DONE]
- Aug-Oct 2019 Adding more English translations to the training data [DONE]
- Oct 2019 Task registrations open [DONE]
- Oct-Dec 2019 Test data annotation [DONE]
- Jun 30 2020 Test data released / Task registrations due [DONE]
- Jul 312020 Run submissions due
- Aug 31 2020 Evaluation results released
- Dec 2020 NTCIR-15 Conference at NII, Tokyo, Japan

Timezone: Japan (UTC+9)

Have questions?

Please contact: dialeval1org@list.waseda.jp

Conditions and Terms

See https://dialeval-1.github.io/dataset/terms