Overview of the NTCIR-15 Dialogue Evaluation (DialEval-1) Task
Zhaohao Zeng, Sosuke Kato, Tetsuya Sakai (Waseda University), Inho Kang (Naver)
dialeval1org@list.waseda.jp
Homepage: http://sakailab.com/dialeval1/

Task as One Sentence

**DQ:** The Dialogue Quality subtask requires the systems to predict distributions of dialogue quality scores.

**ND:** The Nugget Detection subtask requires the systems to predict distributions of nugget types.

**Motivation**

**Why prediction?**
To build good dialogue system, we need good ways to evaluate them.

- Online evaluation:
  - Expensive and does not scale
  - Difficult to compare different systems
  - Not repeatable

**Why distributions?**
People can have diverse views for the same dialogue.

**Evaluation**

**Dialogue Quality subtask**

- **INPUT:** a customer-helpdesk dialogue \(d \in D\)
- **OUTPUT:** an estimated probability distribution \(p\) of dialogue quality score

\[
meanM = \frac{1}{|D|} \sum_{d \in D} M(d)
\]

Compared two distributions over ordinal bins (dialogue quality scores)

**Nugget Detection subtask**

- **INPUT:** a customer-helpdesk dialogue \(d \in D\)
- **OUTPUT:** estimated \(p\)'s over customer nugget types

\[
\text{M(d): measure quantifying how } p \text{ differs from } p^* \text{(see later slide)}
\]

\[
\text{for the Chinese DQ and ND subtasks, - Two BERT-based models outperformed the BiLSTM (Bidirectional Long Short-term Memory) baseline model (BL-lstm) with statistical significance}
\]

**Summary**

For the Chinese DQ and ND subtasks,
- Two BERT-based models outperformed the BiLSTM (Bidirectional Long Short-term Memory) baseline model (BL-lstm) with statistical significance

For the English DQ and ND subtasks,
- None of the models outperformed the BiLSTM baseline.

**Future**

The English translation for all the dialogues will be available to form a fully bilingual dataset at next DialEval.