The E2E Dataset: New Challenges for End-to-End Generation

Jekaterina Novikova, Ondřej Dušek & Verena Rieser

Interaction Lab
Heriot-Watt University
Edinburgh, Scotland, UK

SIGDIAL, Saarbrücken, 16 August 2017

Motivation

- **E2E NLG**: Learning from data **without alignments**
  - just MR + textual reference
- **So far limited** to small, delexicalized datasets
  - e.g. BAGEL, SF Restaurants, SF Hotels, RoboCup
- Our **goal**: replicate **rich dialogue & discourse** phenomena
  - as targeted by earlier rule-based & statistical approaches

The E2E NLG Dataset

- New corpus for NLG in the restaurant domain
- **More challenging** than previous sets
  - More data: **50k** unaligned **MR+ref pairs**
  - More sentences per one MR
  - Longer sentences
- **More diverse** & natural
  - collected by crowdsourcing with **pictorial instructions**
  - mean 8.1 refs per MR


```
name [Loch Fyne],
eatType[restaurant],
food[Japanese],
price[cheap],
kid-friendly[yes]
```

Serving low cost Japanese style cuisine, Loch Fyne caters for everyone, including families with small children.

Loch Fyne is a kid-friendly restaurant serving cheap Japanese food.

E2E Dataset Properties

HERIOT WATT UNIVERSITY

- Lexical richness
  - **higher lexical diversity** (Mean Segmental Token-Type Ratio)
  - higher proportion of **rare words**
- Syntactic richness
  - **more complex** sentences (D-Level)
- Semantic richness
  - crowd workers asked to verbalize *relevant* information
  - requires **content selection**

The E2E NLG Challenge

HERIOT WATT UNIVERSITY

- Get the data & try your system!

  deadline: 31 October

- Data, baseline system & metrics available now
- See more at **http://bit.ly/e2e-nlg**

Thanks

HERIOT WATT UNIVERSITY

- Come see our poster!
- Get the data & take part in the E2E challenge:

  http://bit.ly/e2e-nlg

  deadline: 31 October
- Contact us:

  j.novikova / o.dusek / v.t.rieser @hw.ac.uk