README: the E2E NLG Challenge

The E2E NLG Challenge is a shared task on generating natural language restaurant descriptions from sets of attribute-value pairs. It reports the results of the first shared task on end-to-end (E2E) natural language generation (NLG) for conversational agents. This task aims to assess which NLG approaches can generate better-quality output by learning from a dataset containing lexical richness, syntactic complexity and diverse discourse phenomena. They compared 62 systems submitted by 17 institutions, covering a wide range of approaches, including machine learning architectures – with the majority implementing sequence-to-sequence models (seq2seq) – as well as systems based on grammatical rules and templates.

The aim of this challenge was to generate an utterance from a given meaning representation (MR), which is

1) similar to human-generated reference texts, and
2) highly rated by humans.

The similarity is assessed using standard automated NLG and MT metrics, such as BLEU and METEOR, along with human ratings obtained using a mixture of crowd-sourcing and expert annotation. A suite of novel metrics were also tested to estimate the quality of a generated utterance. The metrics used for automatic evaluation are available here on Github.

Flat MR	NL reference
name[Loch Fyne], eatType[restaurant], food[French], priceRange[less than £20], familyFriendly[yes]	Loch Fyne is a family-friendly restaurant providing wine and cheese at a low cost. Loch Fyne is a French family friendly restaurant catering to a budget of below £20. Loch Fyne is a French restaurant with a family setting and perfect on the wallet.

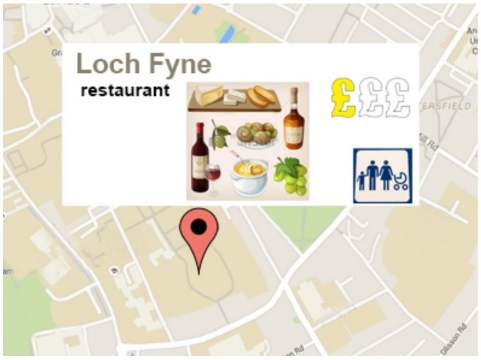

Figure 1. Information about a restaurant (Loch Fyne) in a meaning representation consistent of a list of attribute-value pairs, corresponding natural language reference sentences, and similar information about the restaurant presented pictorially, along with the map location.

Details about the E2E NLG Challenge Dataset:

- The E2E dataset can be downloaded here. A detailed description of the data can be found in the SIGDIAL 2017 paper. A brief summary of the E2E NLG Challenge results is available

in the NLG 2018 paper.

- As shown in Figure 1 above, the dataset consists of attribute-value (AV) pair meaning representations (MRs) and pictorial ones, along with multiple natural language references.
- AV and pictorial MRs correspond to **80%** & **20%** of the dataset respectively. There are fewer pictorial MRs because the corresponding natural language sentences are quite variable.
- The dataset contains **50k** references for **6k** distinct MR's (~**8.27** references/MR).
- The dataset is split into **train, test** & **evaluation set** in the ratio of **82:9:9**.

Baseline & Results of submitted models:

The baseline system for the challenge is TGen (Dusek and Jurcicek, 2016). TGen is a seq2seq model with attention (Bahdanau et al., 2015) with beam search and a reranker to penalize outputs that differ too much from the input MR. As reported on the challenge site at the baseline tab, the TGen baseline scores on the development set are as follows:

TGen on Development Set	
Metric	**Score**
BLEU	0.6925
NIST	8.4781
METEOR	0.4703
ROUGE-L	0.7257
CIDEr	2.3987

The full baseline system outputs can be downloaded for both the development and test sets (one instance per line), from the challenge site, baseline tab. Instructions to run the baseline are provided in the TGen Github repository.

The paper reporting the *Findings of the E2E NLG Challenge* is on Canvas, and at the ACL anthology.