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

Jekaterina Novikova, Ondřej Dušek and Verena Rieser
School of Mathematical and Computer Sciences
Heriot-Watt University, Edinburgh
j.novikova, o.dusek, v.t.rieser@hw.ac.uk

Abstract
This paper describes the E2E data, a new dataset for training end-to-end, data-driven natural language generation systems in the restaurant domain, which is ten times bigger than existing, frequently used datasets in this area. The E2E dataset poses new challenges: (1) its human reference texts show more lexical richness and syntactic variation, including discourse phenomena; (2) generating from this set requires content selection. As such, learning from this dataset promises more natural, varied and less template-like system utterances. We also establish a baseline on this dataset, which illustrates some of the difficulties associated with this data.

1 Introduction
The natural language generation (NLG) component of a spoken dialogue system typically has to be re-developed for every new application domain. Recent end-to-end, data-driven NLG systems, however, promise rapid development of NLG components in new domains: They jointly learn sentence planning and surface realisation from non-aligned data (Dušek andJurčíček, 2015; Wen et al., 2015; Mei et al., 2016; Wen et al., 2016; Sharma et al., 2016; Dušek andJurčíček, 2016a; Lampouras and Vlachos, 2016). These approaches do not require costly semantic alignment between meaning representations (MRs) and the corresponding natural language (NL) reference texts (also referred to as “ground truths” or “targets”), but they are trained on parallel datasets, which can be collected in sufficient quality and quantity using effective crowdsourcing techniques, e.g. (Novikova et al., 2016). So far, end-to-end approaches to NLG are limited to small, delexicalised datasets, e.g. BAGEL (Mairesse et al., 2010), SF Hotels/Restaurants (Wen et al., 2015), or RoboCup (Chen and Mooney, 2008). Therefore, end-to-end methods have not been able to replicate the rich dialogue and discourse phenomena targeted by previous rule-based and statistical approaches for language generation in dialogue, e.g. (Walker et al., 2004; Stent et al., 2004; Demberg and Moore, 2006; Rieser and Lemon, 2009).

In this paper, we describe a new crowdsourced dataset of 50k instances in the restaurant domain (see Section 2). We analyse it following the methodology proposed by Perez-Beltrachini and Gardent (2017) and show that the dataset brings additional challenges, such as open vocabulary, complex syntactic structures and diverse discourse phenomena, as described in Section 3. The data is openly released as part of the E2E NLG challenge.

We establish a baseline on the dataset in Section 4, using one of the previous end-to-end approaches.

2 The E2E Dataset
The data was collected using the CrowdFlower platform and quality-controlled following Novikova et al. (2016). The dataset provides infor-
mation about restaurants and consists of more than 50k combinations of a dialogue-act-based MR and 8.1 references on average, as shown in Table 1. The dataset is split into training, validation and testing sets (in a 76.5-8.5-15 ratio), keeping a similar distribution of MR and reference text lengths and ensuring that MRs in different sets are distinct. Each MR consists of 3–8 attributes (slots), such as name, food or area, and their values. A detailed ontology of all attributes and values is provided in Table 2. Following Novikova et al. (2016), the E2E data was collected using pictures as stimuli (see example in Figure 1), which was shown to elicit significantly more natural, more informative, and better phrased human references than textual MRs.

### 3 Challenges

Following Perez-Beltrachini and Gardent (2017), we describe several different dimensions of our dataset and compare them to the BAGEL and SF Restaurants (SFRest) datasets, which use the same domain.

#### Size:
Table 3 summarises the main descriptive statistics of all three datasets. The E2E dataset is significantly larger than the other sets in terms of instances, unique MRs, and average number of human references per MR (Refs/MR). While having more data with a higher number of references per MR makes the E2E data more attractive for statistical approaches, it is also more challenging than previous sets as it uses a larger number of sentences in NL references (Sents/Ref; up to 6 in our dataset compared to typical 1–2 for other sets) and a larger number of slot-value pairs in MRs (Slots/MR). It also contains sentences of about double the word length (W/Ref) and longer sentences in references (W/Sent).

### Lexical Richness:
We used the Lexical Complexity Analyser (Lu, 2012) to measure various dimensions of lexical richness, as shown in Table 4. We complement the traditional measure of lexical diversity type-token ratio (TTR) with the more robust measure of mean segmental TTR (MSTTR) (Lu, 2012), which divides the corpus into successive segments of a given length and then calculates the average TTR of all segments. The higher the value of MSTTR, the more diverse is the measured text. Table 4 shows our dataset has the highest MSTTR value (0.75) while Bagel has the lowest one (0.41). In addition, we measure lexical sophistication (LS), also known as lexical rareness, which is calculated as the proportion of lexical word types not on the list of 2,000 most frequent words generated from the British National Corpus. Table 4 shows that our dataset contains about 15% more infrequent words compared to the other datasets.

We also investigate the distribution of the top 25 most frequent bigrams and trigrams in our dataset (see Figure 2). The majority of both trigrams (61%) and bigrams (50%) is only used once in the dataset, which creates a challenge to efficiently train on this data. Bigrams used more than once in the dataset have an average frequency of 54.4 (SD = 433.1), and the average frequency of trigrams used more than once is 19.9 (SD = 136.9). For comparison, neither SFRest nor Bagel dataset contains bigrams or trigrams that are only used once. The minimal frequency of bigrams is 27 for Bagel (Mean = 98.2, SD = 86.9) and 76 for SFRest (Mean = 128.4, SD = 50.5), for trigrams the minimal frequency is 24 for Bagel (Mean = 63.5, SD = 54.6) and 43 for SFRest (Mean = 67.3, SD = 18.9). Infrequent words and phrases pose a challenge to efficiently train on this data. Bigrams used more than once in the dataset have an average frequency of 54.4 (SD = 433.1), and the average frequency of trigrams used more than once is 19.9 (SD = 136.9). For comparison, neither SFRest nor Bagel dataset contains bigrams or trigrams that are only used once.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>No. of instances</th>
<th>No. of unique MRs</th>
<th>Refs/MR</th>
<th>Slots/MR</th>
<th>W/Ref</th>
<th>W/Sent</th>
<th>Sents/Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2E</td>
<td>50,602</td>
<td>5,751</td>
<td>8.1 (2–16)</td>
<td>5.43</td>
<td>20.1</td>
<td>14.3</td>
<td>1.5 (1–6)</td>
</tr>
<tr>
<td>SFRest</td>
<td>5,192</td>
<td>1,950</td>
<td>1.82 (1–101)</td>
<td>2.86</td>
<td>8.53</td>
<td>8.53</td>
<td>1.05 (1–4)</td>
</tr>
<tr>
<td>Bagel</td>
<td>404</td>
<td>202</td>
<td>2 (2–2)</td>
<td>5.41</td>
<td>11.54</td>
<td>11.54</td>
<td>1.02 (1–2)</td>
</tr>
</tbody>
</table>

Table 3: Descriptive statistics of linguistic and computational adequacy of datasets.

No. of instances is the total number of instances in the dataset, No. of unique MRs is the number of distinct MRs, Refs/MR is the number of NL references per one MR (average and extremes shown), Slots/MR is the average number of slot-value pairs per MR, W/Ref is the average number of words per MR, W/Sent is the average number of words per single sentence, Sents/Ref is the number of NL sentences per MR (average and extremes shown).

Table 4: Lexical Sophistication (LS) and Mean Segmental Type-Token Ratio (MSTTR).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Tokens</th>
<th>Types</th>
<th>LS</th>
<th>TTR</th>
<th>MSTTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2E</td>
<td>65,710</td>
<td>945</td>
<td>0.57</td>
<td>0.01</td>
<td>0.75</td>
</tr>
<tr>
<td>SFRest</td>
<td>45,791</td>
<td>1,187</td>
<td>0.43</td>
<td>0.03</td>
<td>0.62</td>
</tr>
<tr>
<td>Bagel</td>
<td>1,071</td>
<td>70</td>
<td>0.42</td>
<td>0.04</td>
<td>0.41</td>
</tr>
</tbody>
</table>

Figure 2: Distribution of the top 25 most frequent bigrams and trigrams in our dataset (left: most frequent bigrams, right: most frequent trigrams).

Figure 3: D-Level sentence distribution of the datasets under comparison.

Syntactic Variation and Discourse Phenomena:
We used the D-Level Analyser (Lu, 2009) to evaluate syntactic variation and complexity of human references using the revised D-Level Scale (Lu, 2014). Figure 3 show a similar syntactic variation in all three datasets. Most references in all the datasets are simple sentences (levels 0 and 1), although the proportion of simple texts is the lowest for the E2E NLG dataset (46%) compared to others (47-51%). Examples of simple sentences in our dataset include: “The Vaults is an Indian restaurant”, or “The Loch Fyne is a moderate priced family restaurant”. The majority of our data, however, contains more complex, varied syntactic structures, including phenomena explicitly modelled by early statistical approaches (Stent et al., 2004; Walker et al., 2004). For example, clauses may be joined by a coordinating conjunction (level 2), e.g. “Cocum is a very expensive restaurant but the quality is great”. There are 14% of level-2 sentences in our dataset, comparing to 7-9% in others. Sentences may also contain verbal gerund (-ing) phrases (level 4), either in addition to previously discussed structures or separately, e.g. “The coffee shop Wildwood has fairly priced food, while being in the same vicinity as the Ranch” or “The Vaults is a family-friendly restaurant offering fast food at moderate prices”. Subordinate clauses are marked as level 5, e.g. “If you like Japanese food, try the Vaults”. The highest levels of syntactic complexity involve
sentences containing referring expressions (“The Golden Curry provides Chinese food in the high price range. It is near the Bakers”), non-finite clauses in adjunct position (“Serving cheap English food, as well as having a coffee shop, the Golden Palace has an average customer rating and is located along the riverside”) or sentences with multiple structures from previous levels. All the datasets contain 13-16% of sentences of levels 6 and 7, where Bagel has the lowest proportion (13%) and our dataset the highest (16%).

Content Selection: In contrast to the other datasets, our crowd workers were asked to verbalise all the useful information from the MR and were allowed to skip an attribute value considered unimportant. This feature makes generating text from our dataset more challenging as NLG systems also need to learn which content to realise. In order to measure the extent of this phenomenon, we examined a random sample of 50 MR-reference pairs. An MR-reference pair was considered a fully covered (C) match if all attribute values present in the MR are verbalised in the NL reference. It was marked as “additional” (A) if the reference contains information not present in the MR and as “omitted” (O) if the MR contains information not present in the reference, see Table 5. 40% of our data contains either additional or omitted information. This often concerns the attribute-value pair eatType=restaurant, which is either omitted (“Loch Fyne provides French food near The Rice Boat. It is located in riverside and has a low customer rating”) or added in case eatType is absent from the MR (“Loch Fyne is a low-rating riverside French restaurant near The Rice Boat”).

4 Baseline System Performance

To establish a baseline on the task data, we use TGen (Dušek and Jurčíček, 2016a), one of the recent E2E data-driven systems. TGen is based on sequence-to-sequence modelling with attention (seq2seq) (Bahdanau et al., 2015). In addition to the standard seq2seq model, TGen uses beam search for decoding and a reranker over the top k outputs, penalizing those outputs that do not verbalize all attributes from the input MR. As TGen does not handle unknown vocabulary well, the sparsely occurring string attributes (see Table 2) name and near are delexicalized – replaced with placeholders during generation time (both in input MRs and training sentences).

We evaluated TGen on the development part of the E2E set using several automatic metrics. The results are shown in Table 6. Despite the greater variety of our dataset as shown in Section 3, the BLEU score achieved by TGen is in the same range as scores reached by the same system for BAGEL (0.6276) and SFRest (0.7270). This indicates that the size of our dataset and the increased number of human references per MR helps statistical approaches.

Based on cursory checks, generator outputs seem mostly fluent and relevant to the input MR. For example, our setup was able to generate long, multi-sentence output, including referring expressions and ellipsis, as illustrated by the following example: “Browns Cambridge is a family-friendly coffee shop that serves French food. It has a low customer rating and is located in the riverside area near Crowne Plaza Hotel.” However, TGen requires delexicalization and does not learn content selection, forcing the verbalization of all MR attributes.

TGen is freely available at https://github.com/UFAL-DSG/tgen.

Detailed system training parameters are given in the supplementary material.

To measure the scores, we used slightly adapted versions of the official MT-Eval script (BLEU, NIST) and the COCO Caption (Chen et al., 2015) metrics (METEOR, ROUGE-L, CIDEr). All evaluation scripts used here are available at https://github.com/tuetschek/e2e-metrics.
5 Conclusion

We described the E2E dataset for end-to-end, statistical natural language generation systems. While this dataset is ten times bigger than similar, frequently used datasets, it also poses new challenges given its lexical richness, syntactic complexity and discourse phenomena. Moreover, generating from this set also involves content selection. In contrast to previous datasets, the E2E data is crowdsourced using pictorial stimuli, which was shown to elicit more natural, more informative and better phased human references than textual meaning representations (Novikova et al., 2016). As such, learning from this data promises more natural and varied outputs than previous “template-like” datasets. The dataset is freely available as part of the E2E NLG Shared Task.

In future work, we hope to collect data with further increased complexity, e.g. asking the user to compare, summarise, or recommend restaurants, in order to replicate previous rule-based and statistical approaches, e.g. (Walker et al., 2004; Stent et al., 2004; Demberg and Moore, 2006; Rieser et al., 2014). In addition, we will experiment with collecting NLG data within a dialogue context, following (Dušek and Jurčíček, 2016b), in order to model discourse phenomena across multiple turns.

Acknowledgements

This research received funding from the EPSRC projects DILiGENt (EP/M005429/1) and MaDrIgAL (EP/N017536/1). The Titan Xp used for this research was donated by the NVIDIA Corporation.

References


The E2E Dataset Supplementary Material: Baseline Model Parameters

<table>
<thead>
<tr>
<th>Setting</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam optimizer learning rate</td>
<td>5e-4</td>
</tr>
<tr>
<td>Network cell type</td>
<td>LSTM</td>
</tr>
<tr>
<td>Embedding (+cell) size</td>
<td>50</td>
</tr>
<tr>
<td>Batch size</td>
<td>20</td>
</tr>
<tr>
<td>Encoder length (max. input attribute-value pairs)</td>
<td>10</td>
</tr>
<tr>
<td>Decoder length (max. output tokens)</td>
<td>80</td>
</tr>
<tr>
<td>Max. training epochs</td>
<td>20</td>
</tr>
<tr>
<td>Training instances reserved for validation</td>
<td>2000</td>
</tr>
</tbody>
</table>

Table G: TGen training parameters: main sequence-to-sequence (seq2seq) model with attention.

The training data are tokenized, lowercased, and values of name and near attributes are replaced with placeholders (“X-name”, “X-near”) for training and generation. The generator is trained by minimizing cross entropy in direct token-by-token generation of surface strings. Validation using BLEU on the reserved instances is performed after each epoch. Early stopping is in force: if top 3 BLEU results do not change for 5 epochs, training is finished. Network parameters giving best BLEU in validation are stored and used in the final model. We used 5 different random initializations of the model and selected the one that yielded best validation BLEU for final evaluation.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam optimizer learning rate</td>
<td>1e-3</td>
</tr>
<tr>
<td>Embedding (+cell) size</td>
<td>50</td>
</tr>
<tr>
<td>Batch size</td>
<td>20</td>
</tr>
<tr>
<td>Training epochs</td>
<td>20</td>
</tr>
<tr>
<td>Encoder length (max. input tokens)</td>
<td>80</td>
</tr>
<tr>
<td>Training instances reserved for validation</td>
<td>2000</td>
</tr>
</tbody>
</table>

Table H: TGen training parameters: reranker.

The reranker is trained to classify input MRs based on NL references (to be able to rerank main seq2seq generator outputs based on how well they reflect the input MR). Given that the name and near attributes are delexicalized, the result is a set of 21 classifiers (showing the presence or absence of each possible delexicalized attribute-value pair). Validation on training and validation data is performed after each epoch and best-performing parameters are kept in the end. Validation data has 10 times more importance than training data for validation.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beam size</td>
<td>10</td>
</tr>
<tr>
<td>Reranker misfit penalty</td>
<td>100</td>
</tr>
</tbody>
</table>

Table I: TGen decoding parameters.

The reranking penalty is very high so that outputs that best cover the input MR perfectly are always promoted to the top of the output 10-best list. The correct values for the delexicalized attributes name and near are inserted in a simple postprocessing step.