Computer Speech & Language 59 (2020) 123–156

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge☆

Ondřej Dušek[a,b,*], Jekaterina Novikova[a,c], Verena Rieser[a]

[a] Interaction Lab, Heriot-Watt University, Edinburgh, UK
[b] Charles University, Faculty of Mathematics and Physics, Prague, Czechia
[c] Winterlight Labs, Toronto, Canada

ARTICLE INFO

Article History:
Received 29 April 2019
Accepted 30 June 2019
Available online 3 July 2019

ABSTRACT

This paper provides a comprehensive analysis of the first shared task on End-to-End Natural Language Generation (NLG) and identifies avenues for future research based on the results. This shared task aimed to assess whether recent end-to-end NLG systems can generate more complex output by learning from datasets containing higher lexical richness, syntactic complexity and diverse discourse phenomena. Introducing novel automatic and human metrics, we compare 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. Seq2seq-based systems have demonstrated a great potential for NLG in the challenge. We find that seq2seq systems generally score high in terms of word-overlap metrics and human evaluations of naturalness – with the winning Slug system (Juraska et al., 2018) being seq2seq-based. However, vanilla seq2seq models often fail to correctly express a given meaning representation if they lack a strong semantic control mechanism applied during decoding. Moreover, seq2seq models can be outperformed by hand-engineered systems in terms of overall quality, as well as complexity, length and diversity of outputs. This research has influenced, inspired and motivated a number of recent studies outwith the original competition, which we also summarise as part of this paper.

1. Introduction

This paper provides a comprehensive final report and extended analysis of the first shared task on End-to-End (E2E) Natural Language Generation (NLG), substantially extending previous reports (Novikova and Rieser, 2016; Novikova et al., 2017b; Dušek et al., 2018). In addition to this previous work, we provide a corrected and extended evaluation of the training dataset, as well as a detailed discussion of how current state-of-the-art systems address E2E generation challenges, including semantic accuracy and diversity of outputs, and a comparison of techniques used by the submitted systems with systems outside the competition. We then include a substantially expanded evaluation of the systems using novel automatic metrics, accounting for output complexity, diversity and semantic correctness. In addition, we provide an analysis of system output similarity and confirm that systems using similar techniques, e.g. seq2seq, produce similar outputs. We also provide a detailed error analysis with examples of system outputs. This extended evaluation allows us reach some more in-depth insights about the strength and weaknesses of end-to-end generation systems. Finally, we discuss directions

☆ This paper has been recommended for acceptance by Srinivas Bangalore.

* Corresponding author at: Charles University, Faculty of Mathematics and Physics, Prague, Czechia.
E-mail address: odusek@ufal.mff.cuni.cz (O. Dušek).

https://doi.org/10.1016/j.csl.2019.06.009

for future work with respect to end-to-end generation, as well as NLG evaluation in general. In addition, this paper accompanies a release of all the participating systems' outputs on the test set along with the human ratings collected in the evaluation campaign.

Shared challenges have become an established way of pushing research boundaries in the field of Natural Language Processing, with NLG benchmarking tasks running since 2007 (Belz and Gatt, 2007). These previous shared tasks have demonstrated that large-scale, comparative evaluations are vital for identifying future research challenges in NLG (Belz and Hastie, 2014). The E2E NLG shared task is novel in that it poses new challenges for recent end-to-end, data-driven NLG systems. This type of systems promises rapid development of NLG components in new domains by reducing annotation effort: they jointly learn sentence planning and surface realisation from non-aligned data, e.g. Dušek and Jurčíček (2015), Wen et al. (2015b, 2016), Mei et al. (2016), Sharma et al. (2016), Dušek and Jurčíček (2016a), and Lampouras and Vlachos (2016). As such, these approaches do not require costly semantic alignment between meaning representations (MRs) and the corresponding natural language 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).

At the start of the E2E NLG Challenge, end-to-end approaches to NLG were limited to small, delexicalised datasets, e.g. BAGEL (Mairesse et al., 2010), SF Hotels/Restaurants (Wen et al., 2015b), 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), Mairesse and Walker (2007), and Rieser and Lemon (2009). In this paper, we describe a large-scale shared task based on a new crowdsourced dataset of 50k instances in the restaurant domain (see Section 3). In Section 4, we show that the dataset poses new challenges, such as open vocabulary, complex syntactic structures and diverse discourse phenomena, and that it inspired multiple extensions and further data collection since its original release.

Our shared task aims to assess whether the novel end-to-end NLG systems are able to produce more complex outputs given a larger and richer training dataset. We received 62 system submissions by 17 institutions from 11 countries for the E2E NLG Challenge, with about 1/3 of these submissions coming from industry, as summarised in Section 5. We consider this level of participation an unexpected success, which underlines the timeliness of this task[1] and allows us to reach general conclusions and issue recommendations on the suitability of different methods.

In Section 6, we analyse how the submitted systems address the challenges posed by the dataset and show that the competition inspired further work on our dataset. We evaluate the submitted systems by comparing them to a challenging baseline using automatic evaluation metrics (including novel text-based measures) as well as human evaluation (see Section 7). Note that, while there are other concurrent studies comparing a limited number of end-to-end NLG approaches (Novikova et al., 2017a; Wiseman et al., 2017; Gardent et al., 2017) which emerged during the E2E NLG Challenge, this is the first research to evaluate novel end-to-end generation at scale using human assessment.

Our results in Section 8 show a discrepancy between data-driven seq2seq models versus template- and rule-based systems. While seq2seq models generally score high on word-overlap similarity measures and human rankings of naturalness, manually engineered systems score better than some seq2seq systems in terms of overall quality, as well as diversity and complexity of generated outputs. In Section 9, we conclude by laying out challenges for future shared tasks in this area. We also release a new dataset of 36k system outputs paired with user ratings, which will enable novel research on automatic quality estimation for NLG (Specia et al., 2010; Dušek et al., 2017; Ueffing et al., 2018; Kann et al., 2018; Tian et al., 2018). All data and scripts associated with the challenge, as well as technical descriptions of participating systems are available at the following URL:

http://www.macs.hw.ac.uk/InteractionLab/E2E/

2. Domain and task

In general, the task of NLG is to convert an input MR into a natural language utterance consisting of one or more sentences. In this paper, we focus on the case where an end-to-end data-driven generator is trained from simple pairs of MRs and reference texts, without fine-grained alignments between elements of the MR and words or phrases in the reference texts, as in, e.g. Dušek and Jurčíček (2015) and Wen et al. (2015b). An example pair of a MR and a reference text is shown in Fig. 1. We focus on restaurant recommendations in our experiments, which, previously, have been widely explored in dialogue systems research, e.g. Young et al. (2010), Henderson et al. (2014), and Wen et al. (2017). However, our E2E dataset is substantially bigger and more complex and than previous NLG training datasets for this domain (Mairesse et al., 2010; Wen et al., 2015b) (see Section 4), which allows us to assess whether NLG systems are able to learn to produce more varied and complex utterances given enough training examples (cf. Section 8).

For the input representation, we use a format commonly found in task-oriented domain-specific spoken dialogue systems – unordered sets of *attributes* (slots) and their *values*, e.g. Mairesse et al. (2010), Young et al. (2010), and Liu and Lane (2016).[2] The list of possible attributes used in the MRs in our dataset with example values is shown in Table 1.

[1] Note that, in comparison, the well established Conference in Machine Translation WMT'17 (running since 2006) got 31 institutions submitting to a total of 8 tasks (Bojar et al., 2017a).

[2] Most dialogue systems also include a general intent of the utterance, such as *inform*, *confirm*, or *request* (Young et al., 2010; Wen et al., 2015b; Liu and Lane, 2016). Since our task is focussed on recommendations, this intent would be *recommend/inform* for all our data, and we can therefore disregard it.

MR *name[The Wrestlers], priceRange[cheap], customerRating[1 of 5]*

reference The Wrestlers offers competitive prices, but isn't rated highly by customers.

Fig. 1. Example pair of an MR and a corresponding human-written reference text.

Table 1
Domain ontology of the E2E dataset.

Attribute	**Data type**	Example value
name	verbatim string	*The Eagle,...*
eatType	dictionary	*restaurant, pub,...*
familyFriendly	boolean	*Yes/No*
priceRange	dictionary	*cheap, expensive,...*
food	dictionary	*French, Italian,...*
near	verbatim string	*market square, Cafe Adriatic,...*
area	dictionary	*riverside, city center,...*
customerRating	enumerable	*1 of 5 (low), 4 of 5 (high),...*

3. Data collection procedure

In order to maximise the chances for data-driven end-to-end systems to produce high quality output, we aim to provide training data in sufficient quality and quantity. We turned to crowdsourcing to collect training data in large enough quantities. We used the CrowdFlower platform[3] to recruit workers. Previously, crowdsourcing has mainly been used for evaluation in the NLG community, e.g. Rieser et al. (2014) and Dethlefs et al. (2012). However, recent efforts in corpus creation via crowdsourcing have proven to be successful in related tasks. For example, Zaidan and Callison-Burch (2011) showed that crowdsourcing can result in datasets of comparable quality to those created by professional translators given appropriate quality control methods. Mairesse et al. (2010) demonstrate that crowd workers can produce aligned natural language descriptions from abstract MRs for NLG, a method which also has shown success in related NLP tasks, such as spoken dialogue systems (Wang et al., 2012) or semantic parsing (Wang et al., 2015). More recently, data-driven NLG systems, such as Wen et al. (2015a) and Dušek and Jurčíček (2016), have relied on crowdsourcing for collecting training data.

When crowdsourcing corpora for training NLG systems, i.e. eliciting natural language paraphrases for given MRs from workers, the following main challenges arise:

1. How to ensure the required quality of the collected data?
2. What types of meaning representations can elicit spontaneous, natural and varied data from crowd workers?

In an attempted to address both challenges before collecting the main training dataset for the E2E NLG challenge, we ran a small-scale pre-study published in Novikova et al. (2016). We briefly summarise the results of this study in this section and apply the successful techniques to the whole data set.

For the pre-study, we prepared a subset of 75 distinct MRs, consisting of three, five or eight attributes from our domain (see Table 1) and their corresponding values in order to evaluate MRs with different complexities.[4] We then implemented several automatic validation procedures for filtering the crowdsourced data in order to address (1), see Section 3.1. To address (2), we explored the trade-off between semantic expressiveness of the MR and the quality of crowdsourced utterances elicited for the different semantic representations. In particular, we investigated translating MRs into pictorial representations as used in, e.g. Williams and Young (2007) and Black et al. (2011) for evaluating spoken dialogue systems (see Section 3.2). In the remainder of this section, we first describe the detailed setup used to crowdsource our data (Section 3.3) and then finally evaluate the pre-study by comparing pictorial MRs to text-based MRs used by previous crowdsourcing work (Mairesse et al., 2010; Wang et al., 2012) in Section 3.4.

3.1. Automatic validation measures

We used two simple methods to check the quality of crowd workers on CrowdFlower: first, we only select workers that are likely to be native speakers of English, following Sprouse (2011) and Callison-Burch and Dredze (2010). We use IP addresses to ensure that workers are located in one of three English-speaking countries – Canada, the United Kingdom, or the United States.

[3] The CrowdFlower platform was renamed to FigureEight after our study was completed. See https://www.figure-eight.com/.
[4] The attributes were selected at random, but we excluded MRs that do not contain the attribute *name* as these would not be appropriate for a venue recommendation.

In addition, we included a requirement that "Participants must be native speakers of British or American English" both in the caption of the task listed on CrowdFlower and in the task instructions. Second, we check whether workers spend at least 20s to complete a page of work. This is a standard CrowdFlower option to control the quality of contributions, and it ensures that the contributor is removed from the job if they complete the task too fast.

We also check the quality of the natural language texts produced by crowd workers for a given MR. In particular, we use three JavaScript validators to ensure that the submitted utterances are well-formed English sentences:

1. We check if the ready-to-submit utterance only contains legal characters, i.e. letters, numbers and symbols ", '. : ; £".
2. We check whether the submitted text is not shorter than the required minimal length, which is an approximation of the total number of characters used for all attribute values in a given MR, as calculated by the following equation:

$$min.\ length = \#\ MR\ characters - \#\ MR\ attributes \times 10 \quad (1)$$

Here, *# MR characters* is the total number of characters in the given MR; *# MR attributes* is the number of attributes in the given MR; and 10 is an average length of an attribute name plus two associated square brackets.

3. We check that workers do not submit the same utterance several times.

We ensured by manually checking a small number of initial trial tasks that these automatic validation methods were able to correctly identify and reject 100% of bad submissions.

3.2. Meaning representations: pictures and text

In previous crowdsourcing tasks involving MRs, these were typically presented to workers in a textual form of dialogue acts (Young et al., 2010), such as the following:

$$inform(type = hotel, pricerange = expensive)$$

However, there is a limit in the semantic complexity that crowd workers can handle when using this type of textual/logical descriptions of dialogue acts (Mairesse et al., 2010). Also, Wang et al. (2012) observed that the chosen semantic formalism influences the workers' language, i.e. crowd workers are primed by the words/tokens and ordering used in the MR. Therefore, in contrast to previous work (Mairesse et al., 2010; Wen et al., 2015a; Dušek and Jurčíček, 2016), we explore the usage of different modalities of meaning representation:

- *Textual/logical MRs* appear as a list of comma-separated attribute-value pairs, where attribute values are shown in square brackets after each attribute (see Figs. 1 and 2). The order of attributes is randomised so that crowd workers are not primed by the ordering used in the MRs (Wang et al., 2012).
- *Pictorial MRs* are semi-automatically generated pictures with a combination of icons corresponding to the individual attributes (see Fig. 2). The icons are located on a background showing a map of a city, thus allowing to represent the meaning of the attributes *area* and *near*.

3.3. Data collection setup

We set up the data collection tasks on the CrowdFlower platform, using the automatic checks described in Section 3.1 and using both pictorial and textual MRs as input (see Section 3.2). For this pre-study, we collected 1133 distinct utterances from the 75 distinct/unique MRs we prepared. 744 utterances were elicited using the textual MRs, and 498 utterances were elicited using the pictorial MRs. The data collected in the pre-study are freely available for download.[5] We later used the same CrowdFlower setup to collect the whole E2E NLG dataset (see Section 4).

In terms of financial compensation, crowd workers were paid the standard pay on CrowdFlower, which is $0.02 per page (where each page contained 1 MR). Workers were expected to spend about 20s per page. Participants were allowed to complete up to 20 pages, i.e. create utterances for up to 20 MRs. Mason and Watts (2010) found in their study of financial incentives on Mechanical Turk (counter-intuitively) that increasing the amount of compensation for a particular task does not tend to improve the quality of the results. Furthermore, Callison-Burch and Dredze (2010) observed that there can be an inverse relationship between the amount of payment and the quality of work, because it may be more tempting for crowd workers to cheat on high-paying tasks if they do not have the skills to complete them. Following these findings, we did not increase the payment for our task over the standard level.

[5] See https://github.com/jeknov/INLG_16_submission. The data is not part of the final E2E NLG dataset.

1. *name[Loch Fyne], eatType[restaurant], familyFriendly[yes], priceRange[cheap], food[Japanese]*

2. *name[The Wrestlers], familyFriendly[No], area[riverside], food[Italian], customerRating[5 of 5], priceRange[expensive], near[Cafe Adriatic], eatType[restaurant]*

Fig. 2. Examples of pictorial MRs (left: logical/textual MR, right: corresponding pictorial MR).

3.4. Results and discussion

We analysed the collected natural language reference texts, focussing on textual versus pictorial MRs and their effects on objective measures, such as time taken to collect the data and length of an utterance, and human evaluations of the reference texts collected under the different conditions. Results in full detail can be found in Novikova et al. (2016); here we only summarise the main findings. The data analysis showed that:

- There is no significant difference in the time taken to collect data with pictorial vs. textual MRs.
- The average length of a collected reference text, both in terms of number of characters and number of sentences, depends mainly on the number of attributes associated with the MR, rather than on whether pictures or text were used.
- Compared to textual MRs, pictorial MRs elicit texts that are significantly less similar to the underlying MR in terms of semantic text similarity (Han et al., 2013). We assume that this is because pictorial MRs are less likely to prime the crowd workers in terms of their lexical choices.
- The human evaluation revealed that reference texts produced from pictorial MRs are rated as significantly ($p < 0.01$) more informative than textual MRs. Equally, utterances produced from pictorial MRs were considered to be significantly ($p < 0.001$) more natural and better phrased than utterances collected with textual MRs.[6]

This shows that pictorial MRs have specific benefits for elicitation of NLG data from crowd workers. This may be because the lack of priming by lexical tokens in the MRs leads the crowd workers to producing more spontaneous and natural language, with more variability. As a concrete example of this phenomenon from the collected data, consider the first MR in Fig. 2. The textual version of this MR elicited utterances such as "*Loch Fyne is a family friendly restaurant serving cheap Japanese food.*" whereas the pictorial MR elicited e.g. "*Serving low cost Japanese style cuisine, Loch Fyne caters for everyone, including families with small children.*"

Pictorial stimuli have also been used in other, related NLP tasks, such as crowdsourced evaluations of dialogue systems, e.g. Williams and Young (2007) and Black et al. (2011). Williams and Young (2007), for example, used pictures to set dialogue goals for users (e.g. to find an expensive Italian restaurant in the town centre). However, no analysis was performed regarding the suitability of such representations. This experiment therefore has a bearing on the general issue of human natural language responses to pictorial task stimuli, and shows for example that pictorial task presentations can elicit more natural variability in user inputs to a dialogue system.

Of course, there is a limit in the meaning complexity that pictures can express. We observed that pictorial MRs tend to introduce more noise. In particular, crowd workers tend to omit information, such as *eatType = restaurant*, which is particularly hard to visualise. Finally, producing pictorial MRs is a semi-automatic process, which is expensive to run at large scale.

Based on these findings, we decided to use pictorial MRs to collect 20% of the full dataset and textual MRs for the rest of the data in order to keep noise and production costs low while increasing diversity. To further increase the data quality and diversity, we collected multiple references per MR to help NLG systems deal with potential noise in the data.

[6] Please see Novikova et al. (2016) for a definition of informativeness, naturalness and phrasing.

4. The E2E NLG dataset

Using the procedure described in Section 3, we crowdsourced a large dataset of 50k instances in the restaurant domain (Novikova et al., 2017b). Our dataset is substantially bigger than previous NLG datasets for dialogue in the restaurant domain, i.e. BAGEL (Mairesse et al., 2010) and SF Restaurants (SFRest) (Wen et al., 2015b), which typically only allowed delexicalised data-driven end-to-end approaches (see Section 4.1). In addition, we demonstrate that our data is also more challenging given its lexical richness, syntactic complexity and diverse discourse phenomena. Following an approach suggested by Perez-Beltrachini and Gardent (2017), we describe these different dimensions of our dataset and compare them to the BAGEL and SFRest datasets in Sections 4.2 and 4.3.[7]

To ensure a fair comparison, we analyse both fully lexicalised and delexicalised versions of all datasets. The lexicalised references in all datasets contained full natural language texts including all restaurant names. This is the default form for the E2E set; small postprocessing steps were taken for the other two sets to achieve a compatible format.[8] To obtain the delexicalised versions, we replaced with placeholders (e.g. "X-slot") most slot values from open sets that appear verbatim in the data: restaurant names, area names, addresses, and numbers (see Fig. 3).[9]

Since the E2E and BAGEL datasets contain only restaurant recommendations, i.e. cases where the system is providing information (*inform* dialogue acts), whereas SFRest also includes system questions, confirmations, and greetings, we also created a subset of SFRest dubbed SFRest-inf with only *inform* instances for a fairer comparison.

We processed the datasets using the MorphoDiTa part-of-speech tagger (Straková et al., 2014) to identify tokens, words (as opposed to punctuation tokens) and sentence boundaries. We used the same tagger to preprocess our data for lexical and syntactic complexity analysis.

All code we used for dataset processing and comparison in Sections 4.1–4.3 are freely available for future research under the following URL:

https://github.com/tuetschek/e2e-stats/

The main script downloads all three datasets under comparison, installs and patches different third-party metrics tools, and produces the statistics. The same tools are used to compare system outputs in Section 8.2.

4.1. Size

Table 2 summarises the main size statistics of all three datasets, plus the inform-only portion of SFRest. The E2E dataset is significantly larger than the other sets in terms of the total number of different MRs, the total number of data instances (i.e. MR-reference pairs), and especially in terms of the total amount of text in the human references, which is more than 20 times bigger than the next-biggest SFRest. These differences are even more profound if we consider delexicalisation: almost all MRs in the E2E set are distinct even after delexicalisation, while the number of unique MRs is reduced significantly (by more than half) for the other sets. Delexicalisation also seems to have a less significant effect on the reference texts in the E2E sets than in the other datasets (cf. the number of delexicalised words vs. the total number of words). The high number of instances directly translates to the higher average number of human references per MR, which is 8.27 for the E2E dataset as opposed to less than two for the other sets.[10]

While having more data with a higher number of references per MR makes the E2E data more attractive for statistical approaches and enables learning more robust models, it is also more challenging than previous sets as it contains a larger number of sentences in the human reference texts (up to 6 in our dataset, with an average of 1.54, compared to typically 1–2 for the other sets, which average below 1.1). The sentences themselves are also longer than in the other datasets. This is immediately apparent for SFRest or SFRest-inf, which are up to 40% shorter in terms of words and tokens. BAGEL's sentences are slightly longer than E2E's on average, but this situation is reversed when the sets are delexicalised. In addition, the input MRs in the E2E dataset are more complex than in the other sets: the average number of slot-value pairs in our set is twice that of SFRest (even if only the more complex *inform* dialogue acts are considered), and slightly higher than BAGEL.

The dataset is split into training, validation and test sets (in a 82–9–9 ratio, see Table 3). We ensure that MRs in our test set are all previously unseen, i.e. none of them overlaps with training/development sets, even when restaurant names are removed, unlike the SFRest data (cf. Lampouras and Vlachos, 2016). The test set can be considered adversarial since the MRs contained there are somewhat longer/more complex than those in the training set and the references copy this distribution (cf. Table 3).

[7] The particular versions of the BAGEL and SFRest datasets used for this research are available from http://farm2.user.srcf.net/research/bagel/ and https://www.repository.cam.ac.uk/handle/1810/251304, respectively.

[8] The BAGEL texts are partially delexicalised by default, so we lexicalised them. SFRest texts were detokenised and adverb/plural markers were postprocessed, e.g. "restaurant-s" changed to "restaurants".

[9] This included slot values for *name* and *near* in the E2E dataset, *name, near, phone, address, postcode, count* and *area* in the SFRest dataset, and *name, near, addr, phone, postcode* and *area* in the BAGEL set. For BAGEL, the values *citycentre* and *riverside* were excluded from delexicalisation as they do not always appear verbatim in the data. The delexicalised version of BAGEL is equivalent to how the dataset is distributed by default. SFRest would allow even more delexicalisation in practice – food types and price ranges also appear verbatim in the references. We decided to keep these values lexicalised since they are not from open sets and the two other datasets do not allow for easy delexicalisation in this case.

[10] Note that Refs/MR ratio for the SFRest dataset is skewed: the *goodbye()* MR has up to 101 references, but the average is less than 2 references per MR. This is apparent in the SFRest-inf section, which has a much lower maximum number of references.

E2E	MR	*name[Green Man], food[French], priceRange[more than £30], area[city centre], familyFriendly[no], near[All Bar One]*
	Lex.	Green Man is a French restaurant in the city centre. It is not child friendly and is located near All Bar One. It costs more than thirty pounds.
	Delex.	**X-name** is a french restaurant in the city centre . it is not child friendly and is located near **X-near** . it costs more than thirty pounds .
SFRest	MR	*inform(name='dosa on fillmore', food='indian or indpak', address='1700 fillmore street', phone=4154413672)*
	Lex.	Dosa on fillmore serves indian and indpak food, the address is 1700 fillmore street, and the phone number is 4154413672.
	Delex.	**X-name** serves indian and indpak food , the address is **X-address** , and the phone number is **X-phone** .
BAGEL	MR	*inform(name="Strada", type=placetoeat, eattype=restaurant, area=citycentre, near="The Curry House", near="The Bakers", food=Italian)*
	Lex.	Strada is an Italian restaurant located near The Curry House and The Bakers in the city centre.
	Delex.	**X-name** is an italian restaurant located near **X-near** and **X-near** in the city centre .

Fig. 3. Lexicalised and delexicalised examples from all three compared datasets (with slot placeholders highlighted in delexicalised sentences). Note that the dialogue act for E2E is constant (i.e. "inform") and as such not expressed. SFRest is the only dataset which contains multiple dialogue act types (cf. the SFRest-inf subset).

4.2. Lexical richness

In order to measure various dimensions of lexical richness in the datasets under comparison, we computed statistics on token/unigram, bigram and trigram counts, and we applied the Lexical Complexity Analyser (Lu, 2012), as shown in Table 4. It is clear that our dataset has a much larger vocabulary – 2 × larger than the second largest SFRest, but more than 5 × larger if delexicalised versions of the datasets are considered. This directly translates into the number of distinct lemmas and distinct *n*-grams; the E2E set has almost 10 × more distinct trigrams than SFRest, over 13 × more in the delexicalised versions. While the proportion of *n*-grams only appearing once in the set is slightly lower than in the other datasets, it stays relatively high given the dataset size and narrow domain, and poses a challenging task for end-to-end data-driven approaches.

The traditional measure of lexical diversity is the type-token ratio (TTR):

$$\mathrm{TTR}(\mathrm{text}) = \frac{\#\,\mathrm{distinct\ \ tokens}}{\#\,\mathrm{total\ \ tokens}} \tag{2}$$

However, it is not a good fit in our case when datasets of different sizes in a narrow domain are compared because the values are inversely proportional to the dataset size. Therefore, we complement TTR with the more robust measure of mean segmental TTR (MSTTR) (Lu, 2012), which divides the corpus into successive segments of a given length (50 tokens) 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

Table 2
Overall size statistics for NLG datasets in the restaurant information domain. All statistics for length of MRs and human references are averages (see Section 4.1 for details). Minimum and maximum numbers of references per MR and sentences per reference are shown in brackets below the average. Highest values on each line are typeset in bold.

	E2E	SFRest	SFRest-inf	BAGEL
Total instances	**51,426**	5192	3307	404
Total MRs	**6039**	1914	1845	381
Unique delexicalised MRs	**5963**	733	686	156
Total tokens in all references	**1,166,000**	49,081	37,824	6151
Total words in all references	**1,051,093**	44,338	34,863	5766
Total delex. words in all references	**957,205**	37,758	28,375	4671
Slots per MR	**5.74**	2.63	2.69	5.48
References per MR	**8.27**	1.91	1.65	1.06
	(1–46)	**(1–101)**	(1–33)	(1–2)
Tokens per reference	**22.67**	9.45	11.44	15.23
Words per reference	**20.60**	8.54	10.54	14.27
Delexicalised words per reference	**18.77**	7.27	8.58	11.56
Sentences per reference	**1.54**	1.05	1.07	1.03
	(1–6)	(1–4)	(1–4)	(1–2)
Tokens per sentence	14.68	8.97	10.74	14.82
Words per sentence	13.33	8.11	9.90	13.89
Delexicalised words per sentence	**12.15**	6.90	8.06	11.26

Table 3
Total number of MRs and human references in the E2E dataset sections and their complexity (average numbers of slots per MR and tokens per reference).

E2E data part	MRs	References	Slots/MR	Tokens/Ref
Training set	4862	42,061	5.52	20.27
Development set	547	4672	6.30	24.52
Test set	630	4693	6.91	26.76
Full dataset	6039	51,426	5.74	22.67

higher MSTTR value (0.71) than the other sets (≤ 0.65). The difference is even more profound if we consider delexicalised versions of the sets and inform-only MRs in the SFRest data – 0.66 vs. 0.55 for SFRest-inf and 0.48 for BAGEL.

In addition, we measure *lexical sophistication* (LS2) (Lu, 2012), also known as lexical rareness, which is calculated as the proportion of lexical word types not on the list of 2000 most frequent words generated from the British National Corpus. Table 4 shows that while the E2E is more sophisticated than SFRest, it is slightly less so compared to BAGEL. However, LS2 numbers on the delexicalised sets show that this is mainly caused by lexical slot values – the delexicalised E2E dataset is almost twice as sophisticated as both SFRest and BAGEL.

Following Oraby et al. (2018a) and Jagfeld et al. (2018), we also use Shannon entropy (Manning and Schütze, 2000, p. 61ff.) as a measure of lexical diversity in the texts:

Table 4
Lexical complexity and diversity statistics for NLG datasets in the restaurant information domain. Counts for *n*-grams appearing only once are shown as absolute numbers and proportions of the total number of respective *n*-grams. Highest values on each line are typeset in bold.

Lexicalised sets	**E2E**	**SFRest**	**SFRest-inf**	**BAGEL**
Distinct tokens	**2780**	1249	1157	601
Distinct tokens occurring once	**890**	230	210	205
	(32%)	(18%)	(18%)	(**34%**)
Distinct lemmas	**2369**	1186	1113	583
Distinct bigrams	**30,111**	5729	4969	1601
Distinct bigrams occurring once	**13,794**	2582	2272	904
	(46%)	(45%)	(46%)	(56%)
Distinct trigrams	**100,731**	11,290	9897	2385
Distinct trigrams occurring once	**56,280**	6832	6091	1667
	(56%)	(61%)	(62%)	(70%)
Lexical sophistication (LS2)	0.616	0.428	0.436	**0.655**
Type-token ratio (TTR)	0.002	0.027	0.032	**0.101**
Mean segmental TTR (MSTTR-50)	**0.706**	0.648	0.626	0.654
Unigram entropy	6.821	**7.411**	7.375	6.773
Bigram entropy	10.146	**10.342**	10.202	9.043
Trigram entropy	**12.604**	11.830	11.766	10.159
Bigram next-word conditional entropy	**3.213**	2.714	2.633	2.202
Trigram next-word conditional entropy	**2.448**	1.463	1.552	1.190
Delexicalised sets	**E2E**	**SFRest**	**SFRest-inf**	**BAGEL**
Distinct tokens	**2675**	504	405	183
Distinct tokens occurring once	**871**	116	95	56
	(**33%**)	(23%)	(23%)	(31%)
Distinct lemmas	**2258**	437	357	161
Distinct bigrams	**26,855**	3099	2360	659
Distinct bigrams occurring once	**12,379**	1376	1068	342
	(46%)	(44%)	(45%)	(52%)
Distinct trigrams	**85,736**	6383	5033	1129
Distinct trigrams occurring once	**47,881**	3628	2905	712
	(56%)	(57%)	(58%)	(**63%**)
Lexical sophistication (LS2)	**0.600**	0.323	0.317	0.317
Type-token ratio (TTR)	0.002	0.012	0.013	**0.035**
Mean segmental TTR (MSTTR-50)	**0.663**	0.602	0.553	0.478
Unigram entropy	**6.388**	6.305	5.944	5.294
Bigram entropy	**9.641**	9.083	8.596	7.160
Trigram entropy	**12.122**	10.546	10.173	8.371
Bigram next-word conditional entropy	**3.140**	2.594	2.477	1.780
Trigram next-word conditional entropy	**2.446**	1.414	1.513	1.216

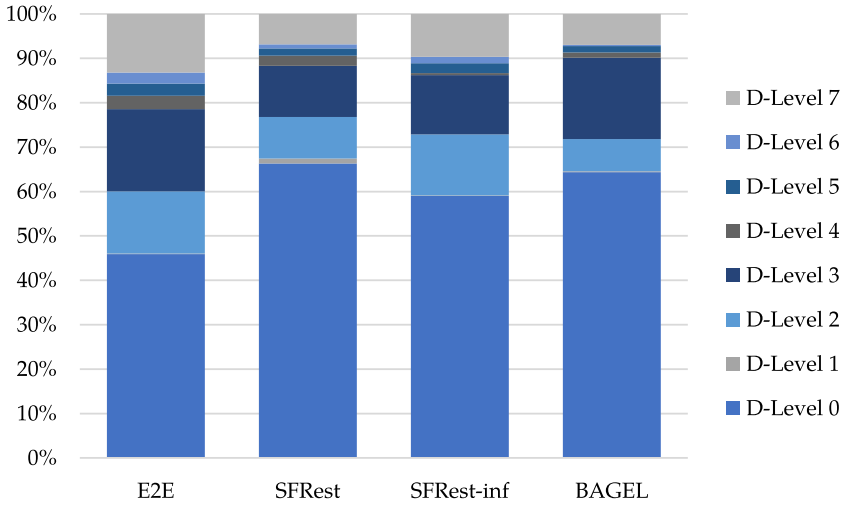


Fig. 4. D-Level sentence distribution of the datasets under comparison.

$$H(\text{text}) = -\sum_{x \in \text{text}} \frac{\text{freq}(x)}{\text{len}(\text{text})} \log_2 \left(\frac{\text{freq}(x)}{\text{len}(\text{text})} \right) \tag{3}$$

Here, x stands for all unique tokens/n-grams, *freq* stands for the number of occurrences in the text, and *len* for the total number of tokens/n-grams in the text. We computed entropy over tokens (unigrams), bigrams and trigrams, as shown in Table 4. We can see that the E2E dataset has slightly lower unigram and bigram entropy than SFRest and higher trigram entropy than any other set. However, when delexicalised, the E2E set shows the highest entropy for any n-gram value. Considering that entropy is a logarithmic measure, the difference is substantial for trigrams – 12.1 vs. the closest 10.5 for SFRest, which amounts to about 2.98 × higher uncertainty.

We further complement Shannon text entropy with n-gram-language-model-style conditional entropy for next-word prediction (Manning and Schütze, 2000, p. 63ff.), given one previous word (bigram) or two previous words (trigram):

$$H_{cond}(\text{text}) = -\sum_{(c,w) \in \text{text}} \frac{\text{freq}(c,w)}{\text{len}(\text{text})} \log_2 \left(\frac{\text{freq}(c,w)}{\text{freq}(c)} \right) \tag{4}$$

Here, (c,w) stands for all unique n-grams in the text, composed of c (context, all tokens but the last one) and w (the last token). Conditional next-word entropy gives an additional, novel measure of diversity and repetitiveness: The more diverse a text is, the less predictable is the next word given previous word(s); on the other hand, the more repetitive the text, the more predictable is the next word given previous word(s). The values for all the datasets are again shown in Table 4, and they demonstrate clearly that E2E data is much more diverse than SFRest or BAGEL. Note also that lexicalisation has a much smaller effect on this measure. In the delexicalised version, the difference against the closest SFRest (2.446 vs. 1.414) indicates about 2.04 × more uncertainty on next-word prediction given two previous words.

4.3. Syntactic complexity

We used the D-Level Analyser (Lu, 2009) to evaluate the syntactic complexity of human references in our data using the revised D-Level scale (Covington et al., 2006). We used the syntactic constituency parser of Collins (1997) to preprocess the sentences for the D-Level Analyser.[11] The D-Level scale has eight levels of syntactic complexity, where levels 0 and 1 include simple or incomplete sentences and higher levels include sentences with more complex structures, e.g. sentences joined by a subordinating conjunction, more than one level of embedding etc. Fig. 4 shows the D-Level distribution in all three datasets.

The largest proportion of the datasets is composed of simple sentences (levels 0 and 1), but the proportion of simple texts is much lower for the E2E NLG dataset (46%) compared to others (59–66%). 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 to NLG (Stent et al., 2004; Walker et al., 2004).[12] For example, clauses may be joined by a

[11] We used the Model 2 variant of the parser as instructed by the D-Level Analyser website at http://www.personal.psu.edu/xxl13/downloads/d-level.html.

[12] Some of the systems in the competition as well as multiple follow-up works are specifically taking advantage of the added complexity in our dataset to produce more varied outputs (see Sections 6.3 and 6.4).

Table 5
Coverage of MR attributes in references as measured manually on a random sample of 50 MR-reference pairs for each dataset. The numbers indicate the absolute number of instances falling into the given category, out of 50.

	E2E	SFRest	BAGEL
Fully covered	30	47	50
Missing content	11	0	0
Additional content	9	3	0

coordinating conjunction (level 2), e.g. "Cocum is a very expensive restaurant *but* the quality is great". There are 14% level-2 sentences in the E2E dataset; BAGEL only has 7% and SFRest 9%, but *inform* MRs in SFRest contain a similar proportion as our set. Level 3 sentences in our domain are mainly those with object-modifying relative clauses, e.g. "There is a pub called Strada *which* serves Italian food." The E2E dataset contains 18% level-3 sentences, similar to BAGEL but more than SFRest's 12% (13% in *inform* MRs). The levels 4–5 are not very frequent in any of the datasets. Sentences may 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 embedded structures from previous levels. As Fig. 4 shows, our dataset has a substantially higher proportion of level-6–7 sentences – 15%, compared to 7% for BAGEL and 8% for SFRest (11% in *inform* MRs).

On average, sentences in the E2E dataset are much more syntactically complex than in the other datasets under comparison: the mean D-Level for E2E data is 2.17, compared to BAGEL's 1.32 and SFRest's 1.25 (1.57 for *inform*-only MRs).

4.4. Attribute coverage

Our crowd workers were asked to verbalise all information from the MR; however, they were not penalised if they skip an attribute (cf. Section 3.4). This feature makes generating text from our dataset more challenging as the NLG systems need to deal with a certain amount of noise, i.e. attributes not being verbalised in the human reference texts. In order to measure the extent of this phenomenon, we examined a random sample of 50 MR-reference pairs in all three datasets under comparison. An MR-reference pair was considered "fully covered" if all attribute values present in the MR are verbalised in the reference. It was marked as "additional content" if the reference contains information not present in the MR, and as "missing content" if the MR contains information not present in the reference.

The results of our sample probe in Table 5 indicate that roughly 40% of our data contains either additional or omitted information. In order to help NLG systems account for this variation, we collected multiple references per MR (also see Table 2).

This variation 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").[13] As discussed in Section 3.4, pictorial MRs might be a possible source of this phenomenon where *eatType=restaurant, eatType=pub*, etc. is difficult to illustrate.

4.5. Following-on datasets

Since the E2E dataset was first published in Novikova et al. (2017b), it inspired multiple extensions (cf. also Section 6.4). The work of Juraska and Walker (2018) adds further automatic annotation of contrast and emphasis to study the style of generation outputs. Oraby et al. (2018a,c) used E2E MRs in combination with the Personage generator (Mairesse and Walker, 2007; 2011) to create a synthetic corpus to examine how neural models can learn various stylistic properties. Reed et al. (2018) then combine Personage-generated data for E2E MRs with the original crowdsourced data and add more supervision specifically to study how neural generators perform various sentence planning operations (sentence aggregation, distribution of content and discourse relations). Balakrishnan et al. (2019) enhance both MRs and references in the E2E set: their enhanced tree-structured MRs include explicit fine-grained dialogue acts (inform, contrast, recommend) automatically obtained from the corresponding references; the enhanced references mark explicitly these dialogue acts as well as which phrase corresponds to which MR attribute. Roberti et al. (2019) present an enhancement of the E2E of a different kind – they include many more restaurant names and food types than present in the original to make the open vocabulary problem more apparent and study it in detail. Finally, the recent

[13] Note that inclusion of this attribute is mainly due to historical reasons, following SFRest and BAGEL.

Table 6
A full list of the primary systems participating in the E2E challenge, with their basic architecture and other properties (list of delexicalised slots, presence of a copy mechanism, control of semantic MR coverage on the output, data augmentation and output diversity techniques). System architectures are coded with colours and symbols: ♥seq2seq, ♦other data-driven, ♣rule-based, ♠template-based.

System	Architecture	Delex. slots	Copy	Semantic control	Data augmentation / diversity
♥**TGen** (Novikova et al., 2017b)	seq2seq (TGen)	*name, near*		MR classification reranking	
♥**Adapt** (Elder et al., 2018)	seq2seq (OpenNMT-py)	none	✓	none	enriching MR by output words
♥**Chen** (Chen, 2018)	seq2seq	none	✓	attention memory	
♥**Gong** (Gong, 2018)	seq2seq (TGen)	*name, near*		MR classification reranking	
♥**Harv** (Gehrmann et al., 2018)	seq2seq	none	✓	coverage penalty reranking	diverse ensembling
♥**NLE** (Agarwal et al., 2018)	char seq2seq (tf-seq2seq)	none		MR classification reranking	
♥**Sheff2** (Chen et al., 2018)	seq2seq	*name, near*		none	
♥**Slug** (Juraska et al., 2018)	seq2seq	*name, near*		slot aligner reranking	using sub-MRs and aligned sentences
♥**Slug-alt** (Juraska et al., 2018) *(late submission) (late submission)*	seq2seq	*name, near*		slot aligner reranking	using only complex training sentences
♥**TNT1** (Oraby et al., 2018b)	seq2seq (TGen)	*name, near*		MR classification reranking	using Personage
♥**TNT2** (Tandon et al., 2018)	seq2seq (TGen)	*name, near*		MR classification reranking	shuffling MRs
♥**TR1** (Schilder et al., 2018)	seq2seq (tf-seq2seq)	*name, near, priceRange, customerRating*		none	
♥**Zhang** (Zhang et al., 2018)	sub-word seq2seq	none		attention regularisation	
♦**Sheff1** (Chen et al., 2018)	linear classifiers + LOLS	*name, near*		2-step prediction with slots	using only references with highest average word frequency
♦**ZHAW1** (Deriu and Cieliebak, 2018)	RNN language model	*name, near*		SC-LSTM (semantic gates), MR classification loss + reranking	first word control
♦**ZHAW2** (Deriu and Cieliebak, 2018)	RNN language model	*name, near*		SC-LSTM (semantic gates)	first word control
♣**DANGNT** (Nguyen and Tran, 2018)	rule-based	all		implied by architecture	
♣**FORGe1** (Mille and Dasiopoulou, 2018)	grammar	all		implied by architecture	
♣**FORGe3** (Mille and Dasiopoulou, 2018)	templates	all		implied by architecture	
♠**TR2** (Schilder et al., 2018)	templates	all		implied by architecture	
♠**TUDA** (Puzikov and Gurevych, 2018)	templates	all		implied by architecture	

Yelp restaurant information dataset of Oraby et al. (2019) is also inspired by E2E, but takes a different approach – collecting large-scale web-based data with automatic annotation.

5. Systems in the competition

The initial idea of the E2E NLG Challenge was first presented in Novikova and Rieser (2016). The interest and active participation in the E2E Challenge has by far outperformed our expectations. We received a total of 62 submitted systems by 17 institutions from 11 countries, with about 1/3 of these submissions coming from industry. In accordance with ethical considerations for NLP shared tasks (Parra Escartín et al., 2017), we allowed researchers to withdraw or anonymise their results after obtaining automatic evaluation metrics results (cf. Section 7.1). Two groups from industry withdrew their submissions and one group asked to be anonymised after obtaining automatic evaluation results. A full list of all the remaining submissions is given in Table A.14 in Appendix A (including their automatic metric scores).

We asked each participating team to identify 1–2 primary systems, which resulted in 20 systems by 14 groups. Each primary system is described in a short technical paper (available on the E2E NLG Challenge website)[14] and was evaluated both by automatic metrics and human judges (see Section 7). We compare the primary systems to a baseline system we provided ourselves (see Section 5.1). A detailed overview of all the primary systems is given in Table 6. In the following, we describe the systems in terms of different architectures; see Sections 5.2–5.5.

[14] http://www.macs.hw.ac.uk/InteractionLab/E2E/

Table 7
TGen performance on the development set (see Section 7.1 for a description of the metrics).

	BLEU	NIST	METEOR	ROUGE-L	CIDEr
TGen (development set)	0.6925	8.4781	0.4703	0.7257	2.3987

5.1. Baseline system

To establish a baseline on the task data, we use TGen (Dušek and Jurčíček, 2016a).[15] TGen is based on the sequence-to-sequence model with attention (seq2seq) (Bahdanau et al., 2015), an encoder-decoder recurrent neural network (RNN) architecture. In addition to the standard seq2seq model with LSTM cells (Hochreiter and Schmidhuber, 1997), TGen uses beam search for decoding and an LSTM-based reranker over the top k outputs, penalising those outputs that do not verbalise all attributes from the input MR. TGen was previously tested on the BAGEL and SFRest datasets, where it reached state-of-the-art performance (Dušek, 2017, p. 88ff.).

As TGen does not handle unknown vocabulary well, the sparsely occurring string attributes (see Table 1) *name* and *near* are delexicalised (see Section 6.1). The main seq2seq model is trained by minimising cross entropy using the Adam algorithm (Kingma and Ba, 2015) in direct token-by-token generation of surface strings; the reranker is trained to detect the presence of all attributes from the input MR.[16] Based on evaluation on the development part of the E2E dataset using automatic metrics (see Table 7), as well as manual cursory checks, TGen appears to be a strong baseline, capable of generating fluent and relevant outputs in most cases.

5.2. Seq2seq-based systems

Systems based on the popular sequence-to-sequence architecture (Sutskever et al., 2014; Bahdanau et al., 2015) represent the biggest group of systems participating in the challenge (12 out of 20 primary systems). All the seq2seq-based systems use beam search, and most of them further enhance the basic seq2seq architecture in a number of ways.

Several systems are built on top of previous systems and toolkits. A number of systems are based on the TGen baseline and aiming to improve it: TNT1 (Oraby et al., 2018b) and TNT2 (Tandon et al., 2018) are using TGen with two different data augmentation techniques (see Section 6.3). Gong (Gong, 2018) trains TGen with fine-tuning by the REINFORCE algorithm (Williams, 1992). Two systems are based on the tf-seq2seq toolkit (Britz et al., 2017): NLE (Agarwal et al., 2018) built a character-to-character seq2seq (using simply characters of the original MR as inputs), TR1 (Smiley et al., 2018) use a regular word-based model. The Adapt system (Elder et al., 2018) is based on OpenNMT-py (Klein et al., 2017). It uses pointer networks (a form of a copy mechanism, Vinyals et al., 2015) and a two-step generation where the first step enriches the input MR for diversity (see Section 6.3).

Several other systems use custom seq2seq implementations. Slug and Slug-alt (Juraska et al., 2018) use an ensemble of two bidirectional LSTM encoders and one convolutional encoder, all paired with an attention LSTM decoder (incl. self-attention). Harv (Gehrmann et al., 2018) use a seq2seq model with multiple additions for MR coverage and diversity (see Sections 6.2 and 6.3). Sheff2's model (Chen et al., 2018), on the other hand, is a vanilla seq2seq setup with LSTM cells. Chen (Chen, 2018) presents a seq2seq model with a custom-tailored input data representation: 2-part input embeddings, which divide into slot name and value token embeddings. Zhang (Zhang et al., 2018) apply a seq2seq model with CAEncoder (Zhang et al., 2017), which adds a second layer over a bidirectional encoder with GRU cells (Cho et al., 2014), summarising both directional encoders.

5.3. Other data-driven systems

Two groups submitted fully trainable systems that are not based on the seq2seq architecture. First, ZHAW1 and ZHAW2 (Deriu and Cieliebak, 2018a) use an RNN language model with semantically conditioned LSTM (SC-LSTM) cells (Wen et al., 2015b) and a 1-hot encoding of input MR slot values. The two system variants differ in the presence of an additional semantic control mechanism (see Section 6.2).

Sheff1 (Chen et al., 2018) is the only non-neural fully data-driven system submitted to the challenge. It is based on imitation learning using linear classifiers (Crammer et al., 2009) in a two-level generation approach, where the classifiers first select the next slot to be realised and then the corresponding word-by-word realisation of that slot (Lampouras and Vlachos, 2016). The classifiers are trained using the Locally Optimal Learning to Search (LOLS) imitation learning framework (Chang et al., 2015), optimising for BLEU, ROUGE-L, and slot error (cf. Section 7.1).

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

[16] We use a learning rate of 0.0005, cell size 50, batch size 20, beam size 10, maximum encoder and decoder lengths 10 and 80, respectively, and up to 20 passes through training data with early stopping. The reranker uses the same parameters, except for a higher learning rate (0.001). See Novikova et al. (2017b) for more details.

5.4. Rule-based systems

There are two rule-based entries in the E2E challenge: first, the DANGNT system (Nguyen and Tran, 2018) uses a two-step rule-based setup, where the first step determines the appropriate phrases to use for a delexicalised sentence; the second step selects the appropriate phrases to lexicalise slot values. Second, the FORGe1 system (Mille and Dasiopoulou, 2018) is a rule-based pipeline using grammars based on the Meaning-Text Theory (Mel'čuk, 1988). It matches the MR to handcrafted per-slot semantic templates, applies aggregation rules to build sentences, and realises the aggregated sentence structures into surface text.

5.5. Template-based systems

Three entries in the E2E challenge are based on traditional template filling. FORGe3 (Mille and Dasiopoulou, 2018) and TR2 (Smiley et al., 2018) take a very similar approach: They mine templates from data by delexicalising slot values. TUDA (Puzikov and Gurevych, 2018), on the other hand, uses templates manually designed by the system authors; the templates are not based on the dataset directly, they are only informed by the data.

6. Addressing the challenges

In this section, we focus on how the competing primary systems address specific challenges posed by the task: vocabulary unseen in training (Section 6.1), control of semantic coverage of the input MR (Section 6.2), and producing diverse outputs (Section 6.3). We also include an overview of alternative approaches to addressing these challenges in Section 6.4.

6.1. Open vocabulary

All systems in the challenge have a way of addressing the open vocabulary in the data. In closed-domain setups, slot values are the usually the only part of data where open vocabulary is present, as e.g. is the case of the *name* and *near* slots in our dataset (see Table 1). The common approach to dealing with open vocabulary in NLG systems is to use delexicalisation (Wen et al., 2015b; see also Section 4), i.e. replacing slot values with placeholders during training and generation time (both in input MRs and training sentences). This approach is indeed one of the principles of template-based systems; accordingly, all template-based entries in the E2E Challenge use full delexicalisation of all slot values (except, perhaps, the binary-valued *familyFriendly*; cf. Table 6). Both rule-based systems also perform full delexicalisation.

The data-driven systems submitted to our challenge mostly opt for partial delexicalisation (see Table 6); the prevailing approach is to delexicalise only the values of the *name* and *near* slots, which allows for very simple pre- and postprocessing since these values usually appear verbatim in the outputs.[17] TR1 is the only data-driven system to use a stronger delexicalisation, which also includes the *priceRange* and *customerRating* slots. Slug and Slug-alt are the only systems to treat values with different morpho-syntactic properties differently (e.g., a value requiring "an" instead of "a" as an article).

Five of the seq2seq systems in the challenge opted for using no delexicalisation and employ alternative ways of addressing open vocabulary: Adapt, Chen and Harv use a copy mechanism (cf. Section 5.2), which allows the system to copy some of the tokens from the input instead of generating them anew. Zhang operates over sub-word units instead of words; these are determined by the byte-pair encoding algorithm and can combine to create previously unseen words (Sennrich et al., 2016). NLE's seq2seq system operates on the character level.

6.2. Semantic control

Most of the participating systems explicitly attempt to realise all slots and thus cope with the noise in the training data (cf. Section 4.4). Full realisation is implied for template and rule-based systems as the templates and rules always relate to specific slots and are chosen based on the slots in the input MR. On the other hand, vanilla seq2seq systems have no way of controlling whether all input slots have been realised. While attention models (Bahdanau et al., 2015) certainly have an influence on this, they are not explicitly trained to attend exactly once to each slot in a vanilla seq2seq setup. Therefore, most seq2seq systems include an additional tool checking the realised parts of the input MR on the output (cf. Table 6).

The most frequent approach among the E2E submissions is a MR classification reranker (Dušek and Jurčíček, 2016a). Here, the generator first produces multiple outputs using beam search, then these are tested for the presence of all slots from the input MR, and deviations from the input are penalised. Apart from the TGen baseline (using a RNN MR classifier, see Section 5.1), this approach is also taken by all systems based on TGen (TNT1, TNT2, Gong) as well as NLE, which uses a logistic regression classifier. Slug and Slug-alt apply a very similar approach: they use a heuristic slot aligner (trained on words and phrases from training data and WordNet) to align outputs to the input MR and penalise for any unaligned slots. Harv do not build a separate classifier or aligner, but use the sum of weights from the attention model (which should not exceed 1 for each token of the input MR) in a penalty term for reranking.

[17] Unlike other slot values, e.g., *area=riverside* might appear as "near the river". Cf. also our remarks on delexicalisation in Section 4 and Footnote [9].

Two seq2seq systems use a direct modification of the attention mechanism instead of reranking at decoding time. Chen includes attention memory (sum of attention distributions so far in the generation process) as an additional input to the attention model. Zhang adds an attention regularisation loss term to the training process, which attempts to keep the sum of weights close to 1 for each input MR token, similarly to Harv's penalty term. Three systems, Adapt, TR1 and Sheff2, do not use any explicit semantic control mechanism.

The non-seq2seq data-driven systems use specific mechanisms to maintain input MR coverage. ZHAW1 and ZHAW2 are based on SC-LSTM cells (Wen et al., 2015b), which include a special gate that keeps track of slots covered so far in the MR. In addition, ZHAW1 uses convolutional MR classifiers to rerank beam search outputs similarly to most seq2seq systems; however, this classification is also used in an additional loss term during training. The Sheff1 system explicitly decides which slot to verbalise next using a separate slot-level classifier, which is optimised to cover the input MR.

6.3. Data augmentation and diversity

The design of the E2E dataset attempts to provide higher text diversity (see Section 4), and several challenge participants made use of this. Others modified the training set simply to achieve better output quality.

Several systems aim at higher output quality by using data augmentation. TNT1 enriches input MRs by prepending them with the corresponding outputs of the Personage generator (Mairesse and Walker, 2007), with the aim to generate more diverse output. TNT2 aims to boost the robustness of the baseline TGen system by re-shuffling slots in the input MRs. Slug uses single sentences from the training data with corresponding aligned parts of the original MR. This increases the amount of training data available and simplifies the task by breaking outputs into smaller (partially) aligned units. Slug-alt, on the other hand, only uses training instances involving complex sentences in an attempt to provide more sophisticated outputs. On the other hand, the system of Sheff1 is trained using only one reference text per training MR; the reference text with the highest average word frequency is selected. While this approach is likely to decrease output diversity, the authors use it to stabilise system training. Harv takes yet another approach in order to both stabilise training and increase diversity, called diverse ensembling (Guzman-Rivera et al., 2012). In an expectation-maximisation fashion, they split the training data instances into subsets that exhibit similar structural properties and style in the natural language references, then train different models on these subsets and deploy them as an ensemble.

Two teams attempt to increase output diversity by directly modifying the generation process. The ZHAW1 and ZHAW2 systems use a first word control mechanism: they generate outputs starting with all (frequent enough) first words from the training set, then select the final output by sampling. ZHAW1 only samples among semantically correct outputs (see Section 6.2). Adapt takes a different approach, adding a preprocessing step before the main generator, which decides upon specific words that should appear on the output. These are then used to enrich the input MR in the main generation step, providing more diversity on the input.

6.4. Systems outside the competition and E2E-inspired work

Solving the challenges outlined above is an ongoing effort addressed by many recent systems. Here we briefly summarise other attempts by systems outside the competition for completeness. Note that many of these approaches are very recent and have been published only after the E2E NLG Challenge ended; some of them are even inspired by the challenge and work with the E2E dataset.

Apart from delexicalisation, which is most often used in the E2E NLG Challenge, various variants of the copy mechanism are the most prominent approach to address open vocabulary in NLG (Wiseman et al., 2017; Lebret et al., 2016; Bao et al., 2018; Kaffee et al., 2018; Wang et al., 2018). Among works using the E2E dataset, Shimorina and Gardent (2018) combine a copy mechanism with delexicalisation. In contrast, Freitag and Roy (2018) use subwords and recast the NLG model as a denoising autoencoder, with shared input and output embeddings (starting from slot values and "filling in" the rest of the sentence on the output). Roberti et al. (2019) explore the use of character-based models and extend the E2E dataset to include a wider variety of restaurant names to showcase their approach.

Attempts at improving semantic accuracy of the generated texts show a wider variety of approaches. Kiddon et al. (2016) use a "checklist model" – the decoder keeps a vector of items used so far during the generation; this is similar to semantic gates of Wen et al. (2015b), which have been used by the ZHAW1 and ZHAW2 systems in our challenge (see Section 6.2). Tran et al. (2017) use a two-level attention model (composed of a standard attention model and a "refiner", an attention-over-attention module) to improve semantic coverage. Nema et al. (2018) combine semantic gating and two-level attention (with attention over slots, slot values, and a combination thereof). The system of Su et al. (2018) and Su and Chen (2018), which is developed on E2E data, explores using multi-level decoder, adding linguistic complexity gradually to maintain output integrity; this is in fact similar to Freitag and Roy's (2018)) approach. A follow-up by Su et al. (2019) explores using multi-objective optimisation for both NLG and language understanding, where the latter serves as regularisation for the former. Other authors working on the E2E dataset explore supplementary inputs for improving semantic correctness: Reed et al. (2018) use an additional supervision signal indicating the desired number of sentences to generate, Freitag and Roy (2018) show that additional unlabelled training data improves semantic coverage in their denoising-autoencoder-based NLG model and Balakrishnan et al. (2019) enhance E2E MRs with more detail on the target linguistic structure, which they then use to constrain decoding.

Since its initial release in Novikova et al. (2017b), the E2E dataset has motivated several authors to explore generating more diverse outputs, mostly with additional supervision signals: The system of Wiseman et al. (2018) learns latent templates (sequences of phrases/slots) while learning to generate, thus allowing more controllability and arguably more diversity of the outputs –

the templates serve as an additional, fine-grained way of specifying the desired shape of the generator output. Reed et al. (2018) explore using the presence of prespecified contrast markers (e.g. *but, although*) as additional supervision, while Juraska and Walker (2018) investigate other stylistic markers and use them to generate sentences of specified type. Oraby et al. (2018a,c) attempt to generate outputs showing different personality traits (represented by the Big Five model) using additional synthetic training data with personality annotation. In an extension of their E2E competing system (Deriu and Cieliebak, 2018a), Deriu and Cieliebak (2018b) add specific syntactic features to the input MRs to control not only the first word of the output, but also first words of all sentences in multi-sentence outputs and specific phrasing for expressing each slot in the MR. Jagfeld et al. (2018) do not add more supervision but compare the diversity produced by word-level and character-level seq2seq models on E2E data, showing better performance of the latter.

Using an in-house restaurant dataset, Nayak et al. (2017) explore using a basic sentence plan specification (slot ordering and sentence grouping) as an additional training signal to increase output diversity. Working in the transport information domain, Dušek and Jurčíček (2016) and Mangrulkar et al. (2018) condition their generators on preceding dialogue context as well as the input MR to obtain greater diversity.

7. Evaluation setup

We evaluated the systems submitted to the E2E challenge using a range of automatic metrics, which we describe in Section 7.1. This includes a novel application of textual measures[18] and a novel usage of standard word-overlap metrics to assess similarity among individual systems. Automatic metrics are popular in NLG (Gkatzia and Mahamood, 2015) because they are cheaper and faster to run than human evaluation. However, sole use of automatic metrics is only sensible if they are known to be sufficiently correlated with human preferences. Recent studies (Novikova et al., 2017a; Reiter, 2018) have demonstrated that this is very often not the case and that automatic metrics only weakly reflect human judgements on system outputs as generated by data-driven NLG. Therefore, we also performed a large-scale crowdsourced human evaluation, as detailed in Section 7.2. For the human evaluation of the 20 primary systems, we address the problem of how to efficiently compare a large number of systems, by:

1. Extending our previous work (Novikova et al., 2018) on rank-based Magnitude Estimation (RankME) and verifying the method at scale;[19]
2. Introducing the data-efficient TrueSkill algorithm (Herbrich et al., 2006; Sakaguchi et al., 2014) to NLG. This allows us to compute an overall ranking by directly comparing the systems, rather than individually assessing them at higher cost, as done by previous NLG challenges (Belz and Hastie, 2014).

7.1. Automatic metrics

We apply two types of automatic metrics: one set assessing the similarity between generated system outputs and natural language references in the corpus using word-overlap-based measures, and another set assessing the complexity and diversity of system outputs using a variety of textual measures.

7.1.1. Word-overlap metrics

For the first set, we selected a range of metrics measuring word-overlap between system output and references, including BLEU and NIST, which are used as standard in machine translation evaluation (Bojar et al., 2017b) and very common in NLG, and several others which were applied in the COCO caption generation challenge (Chen et al., 2015) as well as other NLG experiments (e.g. Lebret et al., 2016; Gardent et al., 2017; Sharma et al., 2016):

BLEU (Papineni et al., 2002) is the harmonic mean of n-gram precisions of the system output with respect to human-authored reference sentences, with $n \in \{1, \ldots, 4\}$, lowered by a brevity penalty if the output is shorter than references. The n-gram precisions are proportions of n-grams in the system output that can be matched in any of the reference sentences. Repeated n-gram matches are clipped to the maximum number of times the n-gram occurs in any single reference.

NIST (Doddington, 2002) is a version of BLEU with higher weighting for less frequent (i.e., more informative) n-grams and a different length penalty. It uses $n \in \{1, \ldots, 5\}$.

METEOR (Lavie and Agarwal, 2007) measures both precision and recall of unigrams by aligning the system output with the individual human references. In addition to exact word matches, it uses fuzzy matching based on stemming and WordNet synonyms. It computes matches against multiple references separately and uses the best-matching one.

ROUGE-L (Lin, 2004) is based on longest common subsequences (LCS) between the system output and the human references, where a common subsequence requires the same words in the same order but allows additional, non-covered words in the middle of either sequence. The final ROUGE-L score is an F-measure based on maximum precision and maximum recall achieved over any of the human references, where precision and recall are computed as length of the LCS divided by the length of the system output and the reference, respectively.

[18] These measures were previously applied by Perez-Beltrachini and Gardent (2017) and this work (see Section 4) to describe datasets, but not for evaluation of NLG outputs.

[19] The original study (Novikova et al., 2018) was limited to comparing 3 similar systems on 100 utterances.

CIDEr (Vedantam et al., 2015) was primarily designed for generated image captions, but is also applicable for NLG in general. CIDEr is computed as the average cosine similarity between the system output and the reference sentences on the level of n-grams, $n \in \{1, \ldots, 4\}$. The importance of the individual n-grams is given by the Term Frequency Inverse Document Frequency (TF-IDF) measure, which weighs an n-gram's frequency in a particular instance against its overall frequency in the whole dataset.

We provided scripts to the challenge participants to run all of these metrics in a simple, easy-to-use way. The scripts are freely available at the following URL:[20] https://github.com/tuetschek/e2e-metrics

In addition to evaluating all NLG systems individually against human-authored reference texts (see Section 8.1), we also apply the same metrics as measures of output similarity among the systems, comparing each system's outputs with all other systems' outputs in place of references (see Section 8.3).

7.1.2. Textual metrics

For the second set of scores, which is intended to measure complexity and diversity in the system outputs, we use the same automatic textual metrics which we used to evaluate the E2E NLG dataset itself (see Sections 4.2 and 4.3), i.e. dimensions of lexical richness, such as lexical sophistication (LS2) and mean segmental token-to-type ratio (MSTTR), and metrics of syntactic complexity, such as levels of the revised D-Level scale.[21] This allows us to both evaluate the diversity and complexity of system outputs and to establish whether the text characteristics are similar to the training and test sets. To focus specifically on the style produced by the individual systems, we delexicalised restaurant names in the system outputs before computing textual metrics scores, since restaurant names could skew some of these metrics as they are mostly composed of infrequent nouns (cf. Section 4.2).

7.2. Human evaluation

The human evaluation was conducted on the 20 primary systems and the baseline using Rank-based Magnitude Estimation (RankME) (Novikova et al., 2018). In an ordinary (i.e. not rank-based) ME task (Bard et al., 1996), subjects provide a relative rating of an experimental sentence to a reference sentence, which is associated with a pre-set/fixed number. If the target sentence appears twice as good as the reference sentence, for instance, subjects are to multiply the reference score by two; if it appears half as good, they should divide it in half, etc. Rank-based ME extends this idea by asking subjects to provide a relative ranking of several target sentences, i.e. not only to the reference sentence, but also to each other.

Rank-based ME was selected for several reasons. First, its use proved to significantly increase the consistency of human ratings, compared to other data collection methods (Novikova et al., 2018). Second, it implies the use of continuous scales, i.e. rating scales without numerical labels and without given end points. Recent studies show that continuous scales allow subjects to give more nuanced judgements (Belz and Kow, 2011; Graham et al., 2013; Bojar et al., 2017a). Third, it explores relative ranking of different systems instead of directly assessing quality of each specific system, which makes it more reliable in the environment of a challenge.

The evaluation was conducted using crowdsourcing based on the CrowdFlower/FigureEight platform. Crowd workers were presented with five randomly selected outputs of different systems corresponding to a single MR, and were asked to evaluate and rank these systems from the best to the worst, ties permitted, using the RankME method.

The final evaluation results were produced using the TrueSkill algorithm (Herbrich et al., 2006; Sakaguchi et al., 2014). TrueSkill produces system rankings by gradually updating a Bayesian estimate of each system's capability according to the "surprisal" of pairwise comparisons of individual system outputs. This way, fewer direct comparisons between systems are needed to establish their overall ranking. In Novikova et al. (2018), we were able to show that TrueSkill is able to reduce the amount of collected human evaluation data without compromising the final ranking results.

Since the performance of some systems may be very similar and a total ordering would not reflect this, we adopt the practice used in machine translation of presenting a partial ordering into significance clusters established by bootstrap resampling (Bojar et al., 2013; 2014; Sakaguchi et al., 2014). The TrueSkill algorithm is run 200 times, producing slightly different rankings each time as pairs of system outputs for comparison are randomly sampled. This way we can determine the range of ranks where each system is placed 95% of the time or more often. Clusters are then formed of systems whose rank ranges overlap.

Traditionally, human evaluation aims to assess the naturalness (fluency, readability) and informativeness (relevance, correctness, adequacy) of an automatically generated output (Gatt and Krahmer, 2017). Naturalness targets the linguistic quality of the NLG system output; informativeness targets relevance or correctness of the output relative to the input MR, showing how well the system reflects the MR content. Recent research often adds a general, overall quality criterion (Wen et al., 2015b; 2015a; Manishina et al., 2016; Novikova et al., 2016; 2017a), or even uses only that (Sharma et al., 2016).

We decided against explicitly evaluating informativeness since our training instances do not always verbalise all MR attributes (cf. Section 4.4). We therefore only collected separate ranks for *quality* and *naturalness*.

Quality: When collecting quality ratings, system outputs were presented to crowd workers together with the corresponding meaning representation, which implies that correctness of the NL utterance relative to the MR should also influence this

[20] The scripts are partially based on COCO caption generation challenge evaluation scripts (https://github.com/tylin/coco-caption).
[21] The script used to install all third-party tools required and run the comparison is available at https://github.com/tuetschek/e2e-stats.

Table 8
Word-overlap metrics scores (see Section 7.1) for all primary systems, plus the average of all metrics' values normalised into the 0–1 range. The list is sorted by the normalised average; any values higher than the corresponding baseline are marked in bold. System architectures are coded with colours and symbols: ♥seq2seq, ♦other data-driven, ♣rule-based, ♠template-based.

System	BLEU	NIST	METEOR	ROUGE-L	CIDEr	norm. avg.
♥**TGen**	0.6593	8.6094	0.4483	0.6850	2.2338	0.5754
♥**Slug**	**0.6619**	**8.6130**	0.4454	0.6772	**2.2615**	0.5744
♥**TNT1**	0.6561	8.5105	**0.4517**	0.6839	2.2183	0.5729
♥**NLE**	0.6534	8.5300	0.4435	0.6829	2.1539	0.5696
♥**TNT2**	0.6502	8.5211	0.4396	**0.6853**	2.1670	0.5688
♥**Harv**	0.6496	8.5268	0.4386	**0.6872**	2.0850	0.5673
♥**Zhang**	0.6545	8.1840	0.4392	**0.7083**	2.1012	0.5661
♥**Gong**	0.6422	8.3453	0.4469	0.6645	**2.2721**	0.5631
♥**TR1**	0.6336	8.1848	0.4322	0.6828	2.1425	0.5563
♦**Sheff1**	0.6015	8.3075	0.4405	0.6778	2.1775	0.5537
♣**DANGNT**	0.5990	7.9277	0.4346	0.6634	2.0783	0.5395
♥**Slug-alt** *(late submission)*	0.6035	8.3954	0.4369	0.5991	2.1019	0.5378
♦**ZHAW2**	0.6004	8.1394	0.4388	0.6119	1.9188	0.5314
♠**TUDA**	0.5657	7.4544	**0.4529**	0.6614	1.8206	0.5215
♦**ZHAW1**	0.5864	8.0212	0.4322	0.5998	1.8173	0.5205
♥**Adapt**	0.5092	7.1954	0.4025	0.5872	1.5039	0.4738
♥**Chen**	0.5859	5.4383	0.3836	0.6714	1.5790	0.4685
♠**FORGe3**	0.4599	7.1092	0.3858	0.5611	1.5586	0.4547
♥**Sheff2**	0.5436	5.7462	0.3561	0.6152	1.4130	0.4462
♠**TR2**	0.4202	6.7686	0.3968	0.5481	1.4389	0.4372
♣**FORGe1**	0.4207	6.5139	0.3685	0.5437	1.3106	0.4231

ranking. The crowd workers were asked: "*How do you judge the overall quality of the utterance in terms of its grammatical correctness, fluency, adequacy and other important factors?*"

Naturalness: When collecting naturalness ratings, system outputs were presented to crowd workers without the corresponding meaning representation. The crowd workers were asked: "*Could the utterance have been produced by a native speaker?*"

Ratings of quality and naturalness were collected separately, i.e. in two individual crowdsourcing tasks. Furthermore, when crowd workers were asked to assess naturalness, the MR was not shown to them since it was not necessary for the task. This setup allows to minimise the correlation between the ratings of naturalness and quality (Novikova et al., 2018; Callison-Burch et al., 2007).

8. Results

In this section, we report on the results of the evaluation of all E2E NLG Challenge primary systems, following the evaluation procedures described in Section 7. We first show the results using automatic metrics: word-overlap-based (Section 8.1) and textual metrics (Section 8.2), as well as automatically computed output similarity between systems (Section 8.3). We then summarise the human evaluation results (Section 8.4), comment on the semantic accuracy of system outputs (Section 8.5) and declare the overall winning system (Section 8.6). Finally, we provide a list of "lessons learnt" in Section 8.7 – observations that we hope will be useful for future NLG system development.

8.1. Word-overlap metrics

Table 8 summarises the system scores for word-overlap metrics (cf. Section 7.1). It is apparent that the TGen baseline is very strong in terms of word-overlap metrics: no primary system is able to beat it in terms of all metrics, or in terms of the normalised metrics' mean – only Slug comes very close. Several other systems manage to beat TGen in one of the metrics but not in others. Note, however, that many secondary system submissions perform better than the primary ones (and the baseline) with respect to word-overlap metrics (see Table A.14 in Appendix A).

Overall, seq2seq-based systems show the best word-based metric values, followed by Sheff1, a data-driven system based on imitation learning. As expected, attempts to increase output diversity by ZHAW1, ZHAW2, Slug-alt and Adapt result in lowered scores by word-overlap-based metrics. Template-based and rule-based systems mostly score at the bottom of the list. The lowest-scoring systems in terms of word-overlap metrics are the ones of Chen and Sheff2, which tend to produce much shorter outputs than other systems (cf. Section 8.2). This most likely resulted in severe brevity penalty.

Finally, it must be noted that the results using automatic metrics are quite different from results obtained in human evaluation (see Section 8.4), which confirms previous findings (Novikova et al., 2017a; Reiter, 2018).

Table 9
Systems sorted according to selected textual metrics (percentage of simple and complex sentences, lexical sophistication LS2, MSTTR-50, average output length in tokens). For comparison, the table also includes the same values for the whole test set (*test set all*) and for a randomly selected subset of the test set, with one reference text per MR (*test set rand*). System architectures are coded with colours and symbols: ▼seq2seq, ♦other data-driven, ♣rule-based, ♠template-based.

% Level0-2		% Level6-7		LS2		MSTTR-50		Avg. length	
▼GONG	82.68	♦SHEFF1	41.27	*test set all*	0.43	*test set rand*	0.62	♠TUDA	31.02
▼TNT2	79.64	♣FORGE1	33.66	*test set rand*	0.36	♠TR2	0.62	♠TR2	27.48
▼SLUG	78.08	▼SLUG-ALT	30.49	▼ADAPT	0.33	▼ADAPT	0.61	♣FORGE1	26.88
▼TNT1	72.18	♦ZHAW1	26.00	♣FORGE1	0.30	♣FORGE1	0.59	♦ZHAW2	26.58
▼ZHANG	70.83	♠TR2	21.07	♠TR2	0.29	♦ZHAW1	0.58	▼TNT1	26.37
♣DANGNT	66.95	♦ZHAW2	19.03	▼HARV	0.27	*test set all*	0.58	♦ZHAW1	26.16
▼TGEN	65.12	♣FORGE3	18.51	▼TNT1	0.26	♦ZHAW2	0.57	▼TNT2	25.49
▼HARV	64.63	*test set rand*	17.46	▼CHEN	0.25	♣FORGE3	0.56	▼GONG	25.41
▼TR1	64.28	▼GONG	16.90	▼NLE	0.25	♠TUDA	0.55	♣DANGNT	24.85
♣FORGE3	62.62	*test set all*	16.48	▼SHEFF2	0.25	♣DANGNT	0.54	▼ADAPT	24.47
▼ADAPT	62.48	▼SLUG	11.39	♦SHEFF1	0.24	▼SLUG-ALT	0.54	▼SLUG-ALT	24.47
♣FORGE1	61.13	▼NLE	11.12	▼TNT2	0.23	▼SLUG	0.52	*test set rand*	24.39
♦ZHAW1	58.91	♠TUDA	10.48	▼TGEN	0.22	▼TNT1	0.52	▼TGEN	24.04
▼NLE	58.24	▼ADAPT	10.28	♣DANGNT	0.21	♦SHEFF1	0.52	*test set all*	23.96
test set rand	58.16	▼TNT1	9.55	♠TUDA	0.21	▼NLE	0.52	▼SLUG	23.76
test set all	57.97	▼TGEN	9.02	▼TR1	0.20	▼TGEN	0.52	♣FORGE3	23.49
♠TUDA	57.66	♣DANGNT	8.91	▼ZHANG	0.20	▼TNT2	0.51	▼NLE	23.40
♠TR2	57.36	▼TR1	8.13	▼SLUG	0.20	▼HARV	0.51	▼HARV	23.22
▼CHEN	54.35	▼HARV	8.12	▼GONG	0.20	▼TR1	0.50	♦SHEFF1	22.75
▼SHEFF2	52.98	▼ZHANG	5.27	♣FORGE3	0.20	▼GONG	0.50	▼TR1	22.43
♦ZHAW2	52.63	▼TNT2	5.22	▼SLUG-ALT	0.19	▼ZHANG	0.47	▼ZHANG	20.71
▼SLUG-ALT	35.12	▼CHEN	4.40	♦ZHAW2	0.17	▼CHEN	0.43	▼SHEFF2	17.18
♦SHEFF1	26.19	▼SHEFF2	2.08	♦ZHAW1	0.17	▼SHEFF2	0.43	▼CHEN	16.32

8.2. Textual metrics

Table 9 summarises results from a range of textual metrics which aim to assess the complexity and diversity of primary system outputs (cf. Section 7.1). In addition, we include a comparison to the human references in the test set in order to assess whether systems are able to replicate characteristics of human-produced data.[22] The results in Table 9 show the following:

- Seq2seq-based system outputs are less syntactically complex on average than outputs of other systems (they produce more D-Level 0–2 sentences and less D-Level 6–7 sentences than other architectures).
- The systems seem to show a relatively high variance in syntactic complexity levels, especially with respect to the higher levels; few systems match the distribution of the training and test data. The differences in D-Level distributions in the outputs are mostly statistically significant (see Fig. A.7 in Appendix A). The only system producing a D-Level distribution *not* significantly different from a random test set reference is FORGE3, which is based on template mining from training data.
  If we use Bhattacharyya distance to compare the D-Level distributions (cf. Fig. A.8 in Appendix A), the greatest distances appear in both extremes. SHEFF1, FORGE1 and SLUG-ALT produce higher-level sentences more frequently and thus show among the most distant from other systems. The GONG system mostly produces level 0–2 sentences, and therefore it appears very distant from other systems as well as the most distant system from human references.
- None of the systems reaches the lexical sophistication of the human-authored test set references. The diversity-attempting seq2seq-based ADAPT system comes very close, followed by the grammar-based FORGE1 and the TR2 system, which is based on template mining from data. Data-driven systems aiming at higher lexical diversity seem to achieve higher sophistication as well; note the lower performance of SLUG-ALT, which aims more at syntactic diversity than lexical. For rule-based systems, lexical sophistication is a direct result of the system authors' decisions.
- In terms of MSTTR, highest scores are achieved by template or rule-based systems and by data-driven systems that explicitly aim at greater output diversity (ZHAW1, ZHAW2, ADAPT, SLUG-ALT). Note that MSTTR is typically higher in systems that tend to produce longer outputs, which includes most rule- and template-based systems. We assume that this is due to MSTTR's fixed 50-token window used to segment utterances.
- Most systems produce outputs similar in length to the test set human references. Outputs of rule- and template-based systems tend to be more verbose than those of data-driven systems. The outputs of ZHANG, SHEFF2 and CHEN are much shorter on average than texts in the dataset, which suggests that these systems might not verbalise all the information contained in the MR (cf. Section 8.5).

Same as for the datasets statistics in Section 4.2, we also computed additional textual measures to assess the diversity/repetitiveness of the generated outputs: number of distinct *n*-grams, Shannon entropy, and conditional next-word entropy; a selection

[22] Note that textual metrics have been computed with restaurant names delexicalised (cf. Section 7.1).

Table 10
Systems sorted according to selected textual diversity metrics (number of distinct tokens, number of distinct trigrams, proportion of unique trigrams, Shannon entropy over tokens (unigrams), bigram next-word conditional entropy). For comparison, the table also includes the same values for the whole test set (*test set all*) and for a randomly selected subset of the test set, with one reference text per MR (*test set rand*). System architectures are coded with colours and symbols: ▼seq2seq, ♦other data-driven, ♣rule-based, ♣template-based.

Distinct tokens		Distinct trigrams		% Unique trigrams		Entropy tokens		Cond. entropy bigrams	
test set all	1079	*test set all*	16797	*test set rand*	69.13	*test set all*	6.40	*test set all*	2.92
test set rand	542	*test set rand*	5166	▼ADAPT	66.61	*test set rand*	6.37	*test set rand*	2.70
▼ADAPT	455	♣TR2	4687	♣TR2	60.44	♣TR2	6.24	♣TR2	2.60
♣TR2	399	▼ADAPT	3567	*test set all*	44.66	▼ADAPT	6.18	▼ADAPT	2.09
♦ZHAW1	136	♦ZHAW1	969	♦ZHAW1	24.97	♣FORGE3	5.74	♣FORGE3	1.66
♣FORGE3	124	♣FORGE3	896	▼HARV	21.88	♦ZHAW1	5.71	▼SLUG-ALT	1.55
♦ZHAW2	102	▼SLUG-ALT	855	▼TNT1	21.34	♦ZHAW2	5.65	▼HARV	1.45
▼HARV	93	▼HARV	777	▼NLE	18.75	▼SLUG-ALT	5.57	♦ZHAW1	1.44
▼TNT1	89	♦ZHAW2	716	♦ZHAW2	18.72	♣FORGE1	5.55	▼TNT2	1.39
♣FORGE1	88	▼TNT1	703	▼SLUG-ALT	18.13	▼HARV	5.50	▼NLE	1.37
▼SLUG-ALT	88	▼TNT2	634	▼CHEN	17.92	♦SHEFF1	5.43	▼TNT1	1.37
▼TNT2	86	▼NLE	608	▼ZHANG	17.81	▼NLE	5.43	♦SHEFF1	1.33
▼TGEN	83	▼TGEN	597	♦SHEFF1	16.44	▼TGEN	5.41	▼TGEN	1.32
▼NLE	81	♦SHEFF1	578	▼SLUG	15.58	▼TNT1	5.37	♦ZHAW2	1.32
▼ZHANG	76	♣FORGE1	549	♣FORGE3	13.50	▼SLUG	5.35	▼TR1	1.30
▼TR1	75	▼ZHANG	511	▼TGEN	13.23	▼TNT2	5.34	♣FORGE1	1.29
▼SLUG	74	▼SLUG	507	▼TNT2	12.93	♣DANGNT	5.29	▼ZHANG	1.26
▼CHEN	73	▼CHEN	480	♣FORGE1	12.39	♣TUDA	5.25	▼CHEN	1.17
♦SHEFF1	72	▼TR1	464	▼TR1	10.78	▼TR1	5.24	▼SLUG	1.13
♣DANGNT	61	♣DANGNT	301	▼GONG	7.30	▼ZHANG	5.21	▼SHEFF2	1.10
▼SHEFF2	59	▼SHEFF2	262	▼SHEFF2	4.96	▼GONG	5.19	♣DANGNT	1.06
▼GONG	58	▼GONG	233	♣DANGNT	0.00	▼CHEN	5.09	▼GONG	0.91
♣TUDA	57	♣TUDA	143	♣TUDA	0.00	▼SHEFF2	4.76	♣TUDA	0.71

of these metrics is shown in Table 10.[23] We compare the outputs against the whole test set (multiple references) and a randomly selected single reference per MR from the test set. The results show the following:

- None of the systems is able to produce as much diversity as is contained in a randomly selected human reference – even the most diverse systems lag behind. ADAPT comes close in vocabulary size, TR2 is the closest system in terms of entropy and next-word conditional entropy.
- In terms of vocabulary, there is a huge gap between the most diverse ADAPT and TR2 systems, and any other system (e.g., the 3rd-ranking ZHAW1 has 3 × smaller vocabulary than TR2, and 2.4 × smaller ratio of unique trigrams).
  TR2 demonstrates that mining templates from the training data can lead to very diverse outputs. FORGE3, which uses the same method, also ranks relatively high on vocabulary size and entropy. The diversity produced by ADAPT's seq2seq model indicates that the preprocessing step enriching the MRs works effectively (cf. Section 6.3).
- All diversity-attempting data-driven systems (ADAPT, ZHAW1, ZHAW2, HARV, TNT1, TNT2, SLUG-ALT) indeed rank better than most systems not incorporating diversity measures, with TNT1 and TNT2 showing lower gains than the rest of the group. However, template-mining-based systems (TR2, FORGE3) produce outputs of similar or higher diversity with no concentrated effort.
- Outputs of seq2seq-based systems which do not explicitly model diversity (e.g. GONG, SHEFF1, TR1, SLUG, CHEN) indeed show lower diversity scores. The rule-based DANGNT system also ranks very low on diversity, and the TUDA system with handcrafted templates is the least diverse of all.

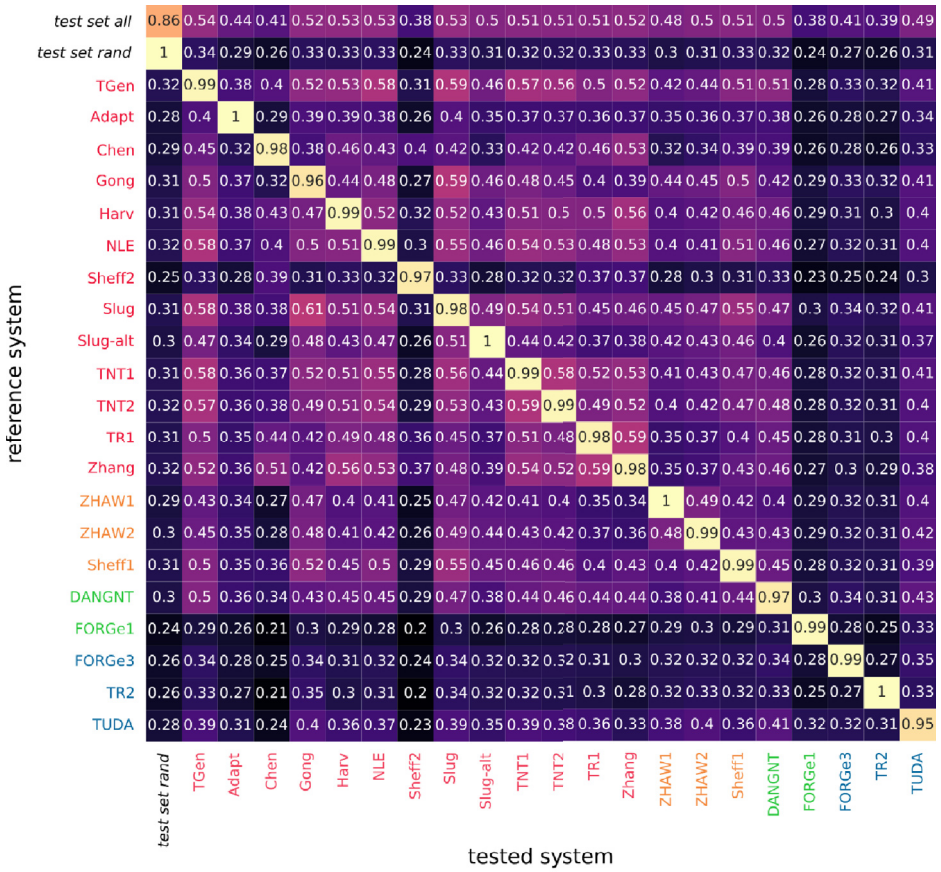
In summary, few systems are able to approach the complexity and diversity shown in human-authored data. Seq2seq-based systems tend to favour simpler sentences than hand-engineered systems unless diversity control is in place. Vanilla seq2seq and handcrafted templates produce the least diverse outputs; highest diversity is achieved by template mining or explicit diversity control mechanisms.

8.3. System output similarity

In order to assess the similarity of outputs produced by the individual systems, we reused the word-overlap-based metrics applied in the challenge (see Section 7.1). We created all possible pairs of systems and computed word-overlap metrics between each of their outputs for every instance in the test set. Same as for textual metrics, restaurant names were delexicalised in the system outputs.[24]

[23] We used system outputs with delexicalised restaurant names for the evaluation, but the lexicalised outputs show the same trends. The values for *n*-gram lengths not displayed in Table 10 also show very similar trends.

[24] Results with fully lexicalised outputs are very similar, the differences are just slightly less profound.



System	Mean
♡TGen	0.48
♡Slug	0.47
♡TNT1	0.46
♡NLE	0.46
♡TNT2	0.46
♡Harv	0.46
♡Zhang	0.45
◇Sheff1	0.44
♡TR1	0.44
♡Gong	0.44
♣DANGNT	0.42
♡Slug-alt	0.42
◇ZHAW2	0.42
◇ZHAW1	0.40
♡Chen	0.40
♠TUDA	0.37
♡Adapt	0.37
♠FORGe3	0.34
♡Sheff2	0.34
test set rand	0.34
♠TR2	0.33
♣FORGe1	0.31

Fig. 5. Similarity of the systems' outputs as measured by automatic metrics (mean of normalised BLEU, NIST, METEOR, ROUGE-L and CIDEr where one system output is used as reference). Systems are sorted by their architecture. For comparison, we also include metrics values against the full test set with multiple references (*test set all*) and against a single-reference randomly sampled subset of the test set (*test set rand*). The table on the right shows mean values of similarity of each system against all other systems (average over columns on the left, excluding the 1st line). System architectures are coded with colours and symbols: ♡seq2seq, ◇other data-driven, ♣rule-based, ♠template-based.

This process resulted in a table for each of the metrics (see Fig. A.6 in Appendix A), with reference systems in rows and tested systems in columns. All five metrics showed a very similar pattern. Fig. 5 therefore summarises the results by taking the average of all normalised metrics (cf. Table 8). For comparison, we also measure similarity of system outputs against the reference texts in the test set, as well as a subset of the test set with a single, randomly sampled reference text per MR.

We can see from Fig. 5 that all the seq2seq-based system outputs are in general most similar to each other; other data-driven systems also show higher similarity amongst each other. The exception to this rule in case of the Chen and Sheff2 systems can be explained by the brevity of their outputs (cf. Sections 8.1 and 8.2). Systems that aim at output diversity (ZHAW1, ZHAW2, Slug-alt and mainly Adapt) also exhibit lowered similarity of their outputs to those of other systems, which might indicate that their outputs are indeed more original. The outputs of rule-based and template-based systems are markedly less similar to other outputs than that of the data-driven systems.

We can also see that most system outputs, especially those of data-driven methods, are much more similar to each other than they are to a single randomly selected human-authored reference text from the test set. This is to be expected since data-driven methods tend to select more frequent phrasing. Some of the system outputs even show a higher similarity to each other than to the closest matching human references from the test set. This is mainly the case for systems with very similar architectures, which often arrive at identical results (e.g. TGen, TNT1 and TNT2).

8.4. Results of human evaluation

The results of human evaluation of quality and naturalness are provided in Table 11. Using the RankME setup described in Section 7.2, we collected 2979 data points of partial system rankings for quality, where one data point corresponds to one MR and ranked outputs of five randomly selected systems (see Table 13 for examples). From these rankings, a set of 29,790 pairwise

Table 11
TrueSkill measurements of quality (left) and naturalness (right) for all primary systems (significance cluster number, TrueSkill value, range of ranks where the system falls in 95% of cases or more, system name). Significance clusters are separated by a dotted line. System architectures are coded with colours and symbols: ♥seq2seq, ♦other data-driven, ♣rule-based, ♠template-based.

Quality					Naturalness				
#	TrueSkill	Rank		System	#	TrueSkill	Rank		System
1	0.300	1	1	♥**SLUG**	1	0.211	1	1	♥**SHEFF2**
2	0.228	2	4	♠**TUDA**	2	0.171	2	3	♥**SLUG**
	0.213	2	5	♥**GONG**		0.154	2	4	♥**CHEN**
	0.184	3	5	♣**DANGNT**		0.126	3	6	♥**HARV**
	0.184	3	6	♥**TGEN**		0.105	4	8	♥**NLE**
	0.136	5	7	♥**SLUG-ALT** *(late)*		0.101	4	8	♥**TGEN**
	0.117	6	8	♦**ZHAW2**		0.091	5	8	♣**DANGNT**
	0.084	7	10	♥**TNT1**		0.077	5	10	♠**TUDA**
	0.065	8	10	♥**TNT2**		0.060	7	11	♥**TNT2**
	0.048	8	12	♥**NLE**		0.046	9	12	♥**GONG**
	0.018	10	13	♦**ZHAW1**		0.027	9	12	♥**TNT1**
	0.014	10	14	♣**FORGE1**		0.027	10	12	♥**ZHANG**
	-0.012	11	14	♦**SHEFF1**	3	-0.053	13	16	♥**TR1**
	-0.012	11	14	♥**HARV**		-0.073	13	17	♥**SLUG-ALT** *(late)*
3	-0.078	15	16	♠**TR2**		-0.077	13	17	♦**SHEFF1**
	-0.083	15	16	♠**FORGE3**		-0.083	13	17	♦**ZHAW2**
4	-0.152	17	19	♥**ADAPT**		-0.104	15	17	♦**ZHAW1**
	-0.185	17	19	♥**TR1**	4	-0.144	18	19	♣**FORGE1**
	-0.186	17	19	♥**ZHANG**		-0.164	18	19	♥**ADAPT**
5	-0.426	20	21	♥**CHEN**	5	-0.243	20	21	♠**TR2**
	-0.457	20	21	♥**SHEFF2**		-0.255	20	21	♠**FORGE3**

output comparisons were produced to be used by the TrueSkill algorithm. This resulted in 1418 pairwise comparisons per system. For naturalness, 4239 data points were collected, which resulted in 42,390 pairwise comparisons, and 2018 comparisons per system. For each of 630 MRs in the test set, 9.5 systems on average (with a maximum of 14) were compared based on both naturalness and quality of their outputs. That is, using TrueSkill, we were able to reduce the number of required system comparisons to more than half. The CrowdFlower task for collecting human evaluation data was running for 235 h and cost USD 314 in total.

We produced the final ranking of all systems for both quality and naturalness using the TrueSkill algorithm with bootstrap resampling as described in Section 7.2. This resulted in clusters of systems with significantly different system rankings for both *naturalness* and *quality*.[25] In both cases, there are clear winning systems (i.e., the 1st cluster only has one member): SHEFF2 for *naturalness* and SLUG for *quality*. The 2nd clusters are quite large for both criteria – they contain 13 and 11 systems, respectively, and they include the baseline TGEN system in both cases.

The results indicate that seq2seq systems dominate in terms of *naturalness* of their outputs, while most systems of other architectures score lower. The bottom cluster is filled with template-based systems. The winning SHEFF2 system is seq2seq-based, and the 2nd cluster mostly includes other seq2seq-based systems. The result also indicates that diversity-attempting systems are penalised in naturalness, i.e. SLUG-ALT, ZHAW1, ZHAW2 placed in the 3rd cluster; ADAPT in the 4th.

The results for *quality*[26] are, however, more mixed in terms of architectures, with none of them clearly prevailing. The 2nd, most populous cluster includes all different architecture types. The winner is the seq2seq-based system SLUG. However, the bottom two clusters are also composed of seq2seq-based systems. This shows the importance of an explicit semantic control mechanism applied at decoding time in seq2seq systems: none of the systems in the bottom two clusters apply such mechanism, whereas all better ranking seq2seq systems do (cf. Section 6.2).[27] Note that this also includes the SHEFF2 system, which scored top for *naturalness*. With the exception of diversity-attempting ADAPT, these systems tend to produce the shortest outputs (see Table 9), which indicates that they are penalised for not realising parts of the input MR too often (cf. Section 8.5).

Finally, we computed the correlation of word-overlap metrics with the human judgements of both quality and naturalness for all the systems. All of the correlations are weak (< 0.2, see Tables A.16 and A.15 in Appendix A), which confirms earlier findings of Novikova et al. (2017a) and explains the discrepancy between system performances in terms of automatic and human evaluation.

[25] Note that TrueSkill provides a relative ranking of a system in terms of its cluster and rank range (cf. Section 7.2), i.e. the numerical scores are not directly interpretable. Other systems in the same cluster are considered to show performance that is not significantly different. In other words: if a system is part of e.g. cluster 2, this system can be considered 2nd best, but it is sharing this position with all other systems in the cluster.

[26] Note that our definition of quality in Section 7.2 also includes semantic completeness and grammaticality.

[27] While the CHEN and ZHANG systems do attempt to model the coverage of the input MR, they do not use explicit beam reranking based on MR coverage.

Table 12
Results of input MR coverage evaluation, with human ratings (left) and using an automatic pattern-matching script (right). Columns legend: *OK* – proportion of outputs covered perfectly, *A* – proportion of outputs with added information, *M* – with missed information, *A+M* with both missed and added information, *SER* – slot/semantic error rate (see Section 8.5). The lists are sorted by the proportion of perfectly covered MRs. System architectures are colour-coded: ♥seq2seq, ♦other data-driven, ♣rule-based, ♠template-based.

Human Ratings					Automatic (pattern matching)					
System	OK	A	M	A+M	System	OK	A	M	A+M	SER
♥SLUG	74%	8%	17%	1%	♠TUDA	100%	0%	0%	0%	0.00%
♥GONG	74%	6%	19%	1%	♦SHEFF1	93%	0%	5%	2%	1.08%
♣DANGNT	74%	9%	17%	0%	♥GONG	92%	4%	2%	2%	1.13%
♠TUDA	74%	19%	7%	0%	♠FORGE1	92%	0%	8%	0%	1.22%
♠TR2	73%	10%	14%	3%	♥SLUG	91%	1%	4%	4%	1.26%
♦SHEFF1	72%	9%	18%	1%	♣DANGNT	88%	0%	12%	0%	1.75%
♥SLUG-ALT	70%	12%	18%	1%	♥TGEN	79%	3%	16%	2%	3.56%
♦ZHAW2	69%	8%	22%	1%	♥SLUG-ALT	78%	4%	9%	9%	3.56%
♥TGEN	69%	7%	23%	1%	♦ZHAW2	76%	3%	20%	1%	3.68%
♣FORGE1	68%	9%	20%	3%	♥TNT1	73%	1%	22%	4%	4.92%
♥TNT1	66%	7%	25%	1%	♥TNT2	71%	1%	28%	1%	6.04%
♥TNT2	62%	9%	28%	1%	♦ZHAW1	70%	3%	25%	2%	5.12%
♦ZHAW1	61%	9%	28%	1%	♠TR2	66%	6%	23%	5%	5.45%
♣FORGE3	60%	10%	29%	1%	♥NLE	63%	3%	24%	10%	6.20%
♥NLE	59%	8%	31%	2%	♥HARV	54%	2%	30%	14%	10.43%
♥HARV	53%	9%	35%	4%	♥ADAPT	50%	3%	36%	10%	12.48%
♥TR1	51%	8%	42%	0%	♥TR1	48%	0%	52%	0%	13.83%
♥ADAPT	51%	12%	33%	4%	♣FORGE3	41%	0%	55%	3%	10.41%
♥ZHANG	43%	8%	49%	0%	♥ZHANG	27%	0%	73%	0%	14.80%
♥CHEN	27%	10%	62%	0%	♥CHEN	11%	0%	88%	1%	23.53%
♥SHEFF2	26%	9%	62%	3%	♥SHEFF2	5%	0%	88%	6%	27.94%

8.5. Error analysis: input MR coverage

In order to clarify the mixed quality evaluation results, we attempted to estimate the number of semantic errors produced by the individual systems in two ways: first, we ran a specific crowdsourced evaluation of systems' coverage of the input MR, where crowd workers were asked to manually annotate missed and added information with respect to the input MR (see Table 12). We did not check for workers' correctness here, and thus we can expect some noise, but the annotations confirm that the systems rated low on quality, most of which also produce very short outputs, also correspond to the ones with the lowest proportion of perfectly covered MRs (CHEN, SHEFF2, ZHANG, TR1 and ADAPT).

Second, semantic errors were computed following Reed et al. (2018), where we implemented a script to estimate the coverage automatically based on regular expression matching.[28] This allowed us to produce an independent estimate of the proportion of outputs with missing or added information (see Table 12). Following Reed et al. (2018), we also computed the slot error rate (SER) using this pattern-matching approach and the following formula:[29]

$$\text{SER} = \frac{\#\text{missed} + \#\text{added} + \#\text{value errors} + \#\text{repetitions}}{\#\text{slots}} \tag{5}$$

Here, *missed* stands for slot values missing from the realisations, *added* denotes additional information not present in the MR (hallucinations), *value errors* denote correctly realised slots with incorrect values (e.g., specifying low price range instead of high), and *repetitions* are values mentioned repeatedly in the outputs; *slots* is the total number of slots/attributes in the test set. SER thus amounts to a proportion of erroneously realised slots. While the absolute numbers for perfectly covered MRs are different from those estimated by humans, they mostly follow the same trend. The SER value is highly correlated with the proportion of perfectly covered MRs.

Both evaluations show that template- and rule-based systems, where MR coverage is implied by the architecture, mostly score high in this regard. However, FORGE3, which uses template mining from training data, scores below average; here, some amount of noise was probably carried over from training data. TUDA, on the other hand, scores high in human ratings and even achieved perfect score by the automatic script (100% perfect coverage), but this is partly given by its low diversity (cf. Section 8.2) – all its templates are probably covered well by the patterns. The results also show that some data-driven

[28] We based the patterns for the individual attribute-value pairs on Reed et al.'s (2018)) script and manually enhanced them using the first 500 instances of the E2E development set.

[29] Note that the coverage and SER values produced by the script is only an estimate as the patterns for a given attribute-value pair will not cover all possible all correct ways to express it. This is different from Wen et al.'s (2015b) computation of SER, where full delexicalisation allowed them to directly count placeholders in the output.

Table 13
Example system outputs with human rankings of quality and a detailed error analysis attempting to interpret the rankings.

#	MR / System name and output		Score
1	name[The Vaults], eatType[restaurant], food[French], priceRange[moderate], area[city centre], familyFriendly[yes], near[Raja Indian Cuisine]		
	▼GONG	The Vaults is a french restaurant in the city centre near Raja Indian Cuisine. It is kids friendly and has a moderate price range.	100
	▼HARV	The Vaults is a moderately priced French restaurant in the city centre near Raja Indian Cuisine.	90
	▲FORGE3	Family-friendly and moderate, The Vaults restaurant is located near Raja Indian Cuisine, in the city centre area.	85
	▼ZHANG	The Vaults is a kid friendly French restaurant in the city centre near Raja Indian Cuisine.	80
	▲SHEFF1	The Vaults is a moderately priced French restaurant near Raja Indian Cuisine in the city centre is.	70
2	name[Cotto], eatType[coffee shop], near[The Bakers]		
	▲TR2	Cotto is a coffee shop located near The Bakers.	100
	▼SLUG-ALT	Cotto is a coffee shop and is located near The Bakers	97
	▼TGEN	Cotto is a coffee shop with a low price range. It is located near The Bakers.	85
	▼GONG	Cotto is a place near The Bakers.	85
	▼SHEFF2	Cotto is a pub near The Bakers.	82
3	name[Clowns], eatType[coffee shop], customer rating[3 out of 5], near[All Bar One]		
	▲SHEFF1	Clowns is a coffee shop near All Bar One with a customer rating of 3 out of 5.	100
	▼ZHANG	Clowns is a coffee shop near All Bar One with a customer rating of 3 out of 5.	100
	▲FORGE3	Clowns is a coffee shop near All Bar One with a rating 3 out of 5.	70
	▲ZHAW2	A coffee shop near All Bar One is Clowns. It has a customer rating of 3 out of 5.	50
	▼SHEFF2	Clowns is a pub near All Bar One.	20
4	name[The Cricketers], eatType[restaurant], customer rating[low], familyFriendly[no], near[Ranch]		
	▼SLUG	The Cricketers is a restaurant near Ranch. It is not family friendly and has a low customer rating.	72
	▼SLUG-ALT	Located near Ranch, The Cricketers is a family-friendly restaurant that is not family-friendly and has a low customer rating.	71
	▼ADAPT	The Cricketers is a non - family - friendly restaurant located near the Ranch. It has a low customer satisfaction rating.	68
	▲FORGE1	The restaurant The Cricketers is near Ranch. The Cricketers, which does not welcome kids, has a low customer rating.	65
	▲TUDA	The Cricketers is a restaurant located near Ranch. It has a low customer rating. It is not family friendly.	56

Each example is shown as ranked for quality by a single crowd worker. The raw RankME scores assigned by the crowd workers are shown; however, note that only relative ranks are used by the TrueSkill algorithm. The outputs within each example are sorted by the score for clarity. For the purpose of error analysis, the rankings may be interpreted in the following way (note that quality rankings include both relevance and fluency): 1. GONG and FORGE3 verbalise all attributes but the latter is less fluent. HARV misses the family-friendliness, ZHANG misses the price information. SHEFF1 misses family-friendliness and is not fluent. 2. TR2 and SLUG-ALT provide perfect and fluent information but SLUG-ALT misses the full stop. GONG does not specify the type of place while TGEN adds irrelevant price range information. SHEFF2 indicates a wrong venue type. 3. SHEFF1 and ZHANG provide perfect and fluent information, FORGE3 is less fluent and ZHAW2 even less than that. SHEFF2 indicates a wrong venue type and misses the customer rating information. 4. SLUG provides a perfect an fluent information. SLUG-ALT is repetitive and ADAPT was probably penalised for lack of detokenisation. FORGE1 and TUDA provide a complete information but are not very fluent.

systems are able to achieve very good coverage (especially SHEFF1, GONG and SLUG, with SER estimates below 1.5%), which confirms the efficacy of their respective semantic control approaches (see Section 6.2). Seq2seq systems without reranking (CHEN, SHEFF2, ZHANG, ADAPT, TR1) score near the bottom of the list in both evaluations.

Both estimates also indicate that missing information is the most common type of problem, added (hallucinated) information occurs less frequently, but still poses a serious problem for utterance generation in task-based dialogue systems.[30] It also appears that both problems are connected – systems hallucinating less frequently tend to miss information more often.

Finally, the scores show that attempts at diversity may hurt semantic accuracy. This is most apparent in ADAPT, the most diverse system with no explicit semantic control mechanism. Other systems with diverse outputs, FORGE3 and HARV, also score lower on coverage. In case of FORGE3, this is due to the above-mentioned noise in the mined templates; HARV's reranking is probably less aggressive than others'. On the other hand, ZHAW1, ZHAW2 and especially SLUG-ALT produce diverse outputs while maintaining good coverage thanks to their very powerful semantic control mechanisms.

8.6. Winning system

We consider the SLUG system (Juraska et al., 2018), a seq2seq-based ensemble system, as the overall winner of this challenge. It received high human ratings for both naturalness and quality, as well as for automatic word-overlap metrics. In contrast to vanilla seq2seq systems, SLUG improves semantic coverage using a heuristic slot aligner in combination with a data augmentation method producing partially aligned examples, which places it among the top-scoring systems in terms of MR coverage (cf. Section 8.5). SLUG's only drawback is the relatively low output diversity; note that repetitive output is considered to be problematic for task-based dialogue systems. A variant of the same system, SLUG-ALT, provides much more output diversity at the cost of slightly lower quality ratings and MR coverage; it maintains higher quality and coverage scores than other diversity-attempting approaches.

[30] Note that this problem appears to be more general since it has also been reported in related fields, including image captioning (Rohrbach et al., 2018), machine translation (Koehn and Knowles, 2017; Lee et al., 2019), and question answering (Feng et al., 2018).

While the SHEFF2 system (Chen et al., 2018), a vanilla seq2seq setup, won in terms of naturalness, it often does not realise all parts of the input MR, which severely affected its quality rating – it placed in the last cluster, ranked 20th–21st out of 21. SHEFF2's outputs also rank very low on complexity and diversity.

Furthermore, the TGEN baseline system turned out hard to beat. It ranked highest on average in word-overlap-based automatic metrics and placed in the 2nd cluster in both quality and naturalness (ranks 3–6 and 4–8 out of 21, respectively). TGEN also fared well (albeit not perfectly) in MR coverage evaluations. On the other hand, TGEN only scored in the middle of the pack on output diversity.

8.7. Lessons learnt and future directions

We attempt to formulate some high-level "lessons learnt" for developing future data-driven NLG systems based on the above results, while we acknowledge that our data is limited to a single domain, and that comparisons are not strictly controlled, i.e. models vary in more than one aspect.

- *Semantic control*: For seq2seq-based systems, a strong semantic control of the generated content seems crucial – beam reranking based on MR classification or heuristic alignments appears to work well while attention-only models perform poorly on our data. Correct semantics is regarded by users as more important than fluency (Reiter and Belz, 2009) and should be prioritised when training the models (cf. also Reiter, 2019).
- *Open vocabulary*: For limited domains such as ours, delexicalisation of open-set attributes still seem to be the best approach. However, the systems of HARV and NLE show character-level models and copy mechanisms are viable alternatives. We believe that the low results of CHEN, ZHANG and ADAPT are due to inferior semantic control, not open-vocabulary handling.
- *Complexity and diversity*: In general, hand-engineered systems seem to outperform neural systems in terms of output diversity and complexity (see Section 8.2); the most diverse outputs are produced by systems using templates mined from training data and data-driven systems with explicit diversity mechanisms.
  Vanilla seq2seq-based systems produce the least diverse outputs: they are essentially probabilistic language models, which tend to settle for the most frequent phrasing, thus penalising length and favouring high-frequency word sequences. Diversity in seq2seq models can be improved by data selection (SLUG-ALT), diverse ensembling (HARV) or sampling from the generated beam (Wen et al., 2015b). In contrast, hand-engineered system authors can control the output complexity and diversity directly: here, TUDA's outputs are very repetitive as its set of handcrafted templates is small, while FORGE3 and TR2 with templates mined from data produce some of the most diverse outputs.
  In general, any systems attempting output diversity need to impose strong semantic control mechanisms to maintain MR coverage.
- *Best method suggestion*: Rule-based methods work quite well for limited domains, such as ours. Low-effort handcrafting (as in TUDA) may lead to correct but repetitive outputs. Seq2seq models with semantic reranking emerge as the best data-driven option, in combination with controlling for diversity and using copy mechanisms to minimise preprocessing.

9. Conclusion

This paper presents the findings of the first shared task on End-to-End Natural Language Generation for Spoken Dialogue Systems. The aim of this challenge was to assess the capabilities of recent end-to-end, fully data-driven NLG systems, which can be trained from pairs of input meaning representations and corresponding texts, without the need for fine-grained semantic alignments. In addition to attracting many participants, the challenge has substantially shaped current NLG research, as it has influenced, inspired and motivated a number of recent studies outwith the original competition.

As part of this challenge, we have created a novel dataset for NLG benchmarking in the restaurant information domain, which is an order-of-magnitude bigger than any previous publicly available dataset for task-oriented NLG and has already been used and extended by multiple follow-up works since its original release. We also provided one of the previous state-of-the art seq2seq-based NLG systems, TGEN (Dušek and Jurčíček, 2016a), as a baseline for comparison. The challenge received 62 system submissions by 17 different participating institutions. The systems submitted ranged from complex seq2seq-based setups with different additions to the architecture, over other data-driven methods and rule-based systems, to simple template-based ones. We evaluated all the entries in terms of five different automatic metrics. 20 primary submissions (as identified by the participants) were further evaluated using a novel, crowdsourced evaluation setup. We also include a novel comparison of systems in terms of automatic textual metrics aimed to assess output complexity and diversity. Our evaluation lets us include several general recommendations for future NLG system development.

In general, seq2seq-based systems produce very similar outputs (as measured by word-overlap, cf. Section 8.3), despite their different implementations. Seq2seq models tend to score high on word-overlap metrics and human evaluations of naturalness, while the scores for other data-driven, rule-based and template-based systems are lower. However, these other types of systems often score better in human evaluations of the overall quality. While the winning SLUG system is seq2seq-based, the results also demonstrated possible pitfalls of using seq2seq models:

1. Vanilla seq2seq models tend to produce short outputs of low diversity and syntactic complexity. Low diversity is especially problematic since it causes repetitive outputs in spoken dialogue systems.

2. Applying a strong semantic control mechanism during decoding is crucial to preserve the input meaning. The most common semantic mistake for systems is to miss out information. However, added information (hallucinations) is also closely linked. Both type of errors can have severe consequences for task-based dialogue systems, depending on the application domain.
3. Addressing these issues is challenging: attempts to improve diversity can often result in lowered semantic accuracy and/or output naturalness.[31]

In comparison, hand-engineered systems tend to produce more complex and diverse outputs and are able to reach high overall quality, but are mostly rated low on naturalness. Note that similar findings have been reported by Wiseman et al. (2017) for data-to-document generation. This raises the general question regarding efficiency, costs, and performance of purely data-driven versus carefully hand-engineered NLG systems.

To facilitate further research in this domain, we have made the following data and tools freely available for download:

- The E2E NLG training dataset (including test set with human references),
- A set of word-overlap-based metrics and scripts for running further textual metrics used for automatic evaluation in the challenge,
- Outputs of the baseline TGen system for the development set,
- Outputs for the test set produced by the baseline and all participating systems,
- the corresponding RankME ratings for quality and naturalness collected in the human evaluation campaign.

All can be accessed under the following URL:

http://www.macs.hw.ac.uk/InteractionLab/E2E/

In future work, we aim to investigate additional evaluation methods for NLG systems, such as post-edits (Sripada et al., 2005), or extrinsic evaluation, such as NLG's contribution to task success, e.g. Rieser et al. (2014) and Gkatzia et al. (2016). We also intend to continue our work on automatic quality estimation for NLG (Dušek et al., 2017), where the large amount of data obtained in this challenge allows a wider range of experiments than previously possible.

Acknowledgements

This research received funding from the EPSRC projects DILiGENt (EP/M005429/1) and MaDrIgAL (EP/N017536/1) and Charles University project PRIMUS/19/SCI/10. The Titan Xp used for this research was donated by the NVIDIA Corporation. The authors would like to thank Lena Reed and Shereen Oraby for help with computing the slot error rate. We would also like to thank the anonymous reviewers for providing exceptionally helpful comments and Prof. Ehud Reiter, whose blog[32] inspired some of this research.

[31] This finding is in line with recent follow-up works to the challenge (Oraby et al., 2018a; 2019; Balakrishnan et al., 2019), which suggests that explicit style supervision is needed to produce both diverse and accurate outputs.

[32] https://ehudreiter.com/

Appendix A. Detailed results

Table A.14
Full list of E2E challenge submissions with automatic metric scores (primary systems are indicated in the "P?" column; the column "n. avg." shows an average of all metrics normalised into the 0–1 range, cf. Table 8).

Submitter	System name	P?	BLEU	NIST	METEOR	ROUGE-L	CIDEr	n. avg.
Heriot-Watt Uni	TGen	✓	0.6593	8.6094	0.4483	0.6850	2.2338	0.5754
B. Zhang, Xiamen Uni	Zhang	✓	0.6545	8.1840	0.4392	0.7083	2.1012	0.5661
S. Chen, Harbin Inst of Tech	*abstract beam1*		0.5854	5.4691	0.3977	0.6747	1.6391	0.4737
	abstract beam2		0.5916	5.9477	0.3974	0.6701	1.6513	0.4838
	abstract beam3		0.6150	6.8029	0.4068	0.6750	1.7870	0.5112
	abstract greedy		0.6635	8.3977	0.4312	0.6909	2.0788	0.5666
	non-abstract beam2		0.5860	6.1602	0.3833	0.6619	1.6133	0.4817
	non-abstract beam3		0.6088	6.9790	0.3899	0.6628	1.7015	0.5059
	Chen	✓	0.5859	5.4383	0.3836	0.6714	1.5790	0.4685
Zurich Uni of Applied Sciences	*base*		0.6544	8.3391	0.4448	0.6783	2.1438	0.5652
	ZHAW1	✓	0.5864	8.0212	0.4322	0.5998	1.8173	0.5205
	ZHAW2	✓	0.6004	8.1394	0.4388	0.6119	1.9188	0.5314
	FORGe1	✓	0.4207	6.5139	0.3685	0.5437	1.3106	0.4231
Pompeu Fabra Uni	*2*		0.4113	6.3293	0.3686	0.5593	1.2467	0.4194
	FORGe3	✓	0.4599	7.1092	0.3858	0.5611	1.5586	0.4547
Sheffield NLP	Sheff1	✓	0.6015	8.3075	0.4405	0.6778	2.1775	0.5537
	primary1 var2		0.6233	8.1751	0.4378	0.6887	2.2840	0.5591
	primary1 var3		0.5690	8.0382	0.4202	0.6348	2.0956	0.5275
	primary1 var4		0.5799	7.9163	0.4310	0.6670	2.0691	0.5353
	Sheff2	✓	0.5436	5.7462	0.3561	0.6152	1.4130	0.4462
	primary2 var2		0.5356	7.8373	0.3831	0.5513	1.5825	0.4824
HarvardNLP & Adapt	*support 1*		0.6581	8.5719	0.4409	0.6893	2.1065	0.5712
	support 2		0.6618	8.6025	0.4571	0.7038	2.3371	0.5833
	support 3		0.6737	8.6061	0.4523	0.7084	2.3056	0.5851
	Harv	✓	0.6496	8.5268	0.4386	0.6872	2.0850	0.5673
H. Gong, Harbin Inst of Tech	Gong	✓	0.6422	8.3453	0.4469	0.6645	2.2721	0.5631
	1		0.6396	8.3111	0.4466	0.6620	2.2272	0.5604
	3		0.6395	8.3127	0.4457	0.6628	2.2442	0.5607
	4		0.6395	8.3127	0.4457	0.6628	2.2442	0.5607
Adapt Centre	Adapt	✓	0.5092	7.1954	0.4025	0.5872	1.5039	0.4738
	temperature 0.9		0.5573	7.7013	0.4154	0.6130	1.8110	0.5074
	temperature 1.0		0.5265	7.3991	0.4095	0.5992	1.6488	0.4880
< anonymous >	*combined*		0.2921	4.7690	0.2515	0.4361	0.6674	0.3047
	primary		0.4723	6.1938	0.3170	0.5616	1.2127	0.4183
Naver Labs Europe	NLE	✓	0.6534	8.5300	0.4435	0.6829	2.1539	0.5696
	second		0.6669	8.5388	0.4484	0.6991	2.2239	0.5781
	third		0.6676	8.5416	0.4485	0.6991	2.2276	0.5784
UC Santa Cruz – Slug2Slug	Slug	✓	0.6619	8.6130	0.4454	0.6772	2.2615	0.5744
	Slug-alt *(late)*	✓	0.6035	8.3954	0.4369	0.5991	2.1019	0.5378
Thomson Reuters NLG	*1 model 11 post*		0.6536	8.3293	0.4550	0.6805	2.1050	0.5665
	2 model 13 post		0.6562	8.3942	0.4571	0.6876	2.1706	0.5715
	3 beam 5 model 11 post		0.6805	8.7777	0.4462	0.6928	2.3195	0.5858
	4 beam 5 model 13 post		0.6742	8.6590	0.4499	0.6983	2.3018	0.5837
	5 submission 6		0.6208	8.0632	0.4417	0.6692	2.1127	0.5499
	6 submission 4 beam		0.6201	8.0938	0.4419	0.6740	2.1251	0.5516
	7 submission 4		0.6182	8.0616	0.4417	0.6729	2.0783	0.5494
	8 train only		0.4111	6.7541	0.3970	0.5435	1.4096	0.4336
	TR1	✓	0.6336	8.1848	0.4322	0.6828	2.1425	0.5563
	TR2	✓	0.4202	6.7686	0.3968	0.5481	1.4389	0.4372
UC Santa Cruz – TNT NLG	TNT1	✓	0.6561	8.5105	0.4517	0.6839	2.2183	0.5729
	sys1 model1		0.6476	8.4301	0.4508	0.6795	2.1233	0.5666
	TNT2	✓	0.6502	8.5211	0.4396	0.6853	2.1670	0.5688
	sys2 model1		0.6606	8.6223	0.4439	0.6772	2.1997	0.5728
	sys2 model2		0.6563	8.5482	0.4482	0.6835	2.1953	0.5725
	sys2 model3		0.3681	6.6004	0.3846	0.5259	1.5205	0.4181
VNU-HCM Uni of IT	DANGNT	✓	0.5990	7.9277	0.4346	0.6634	2.0783	0.5395
Tech Uni Darmstadt	TUDA	✓	0.5657	7.4544	0.4529	0.6614	1.8206	0.5215

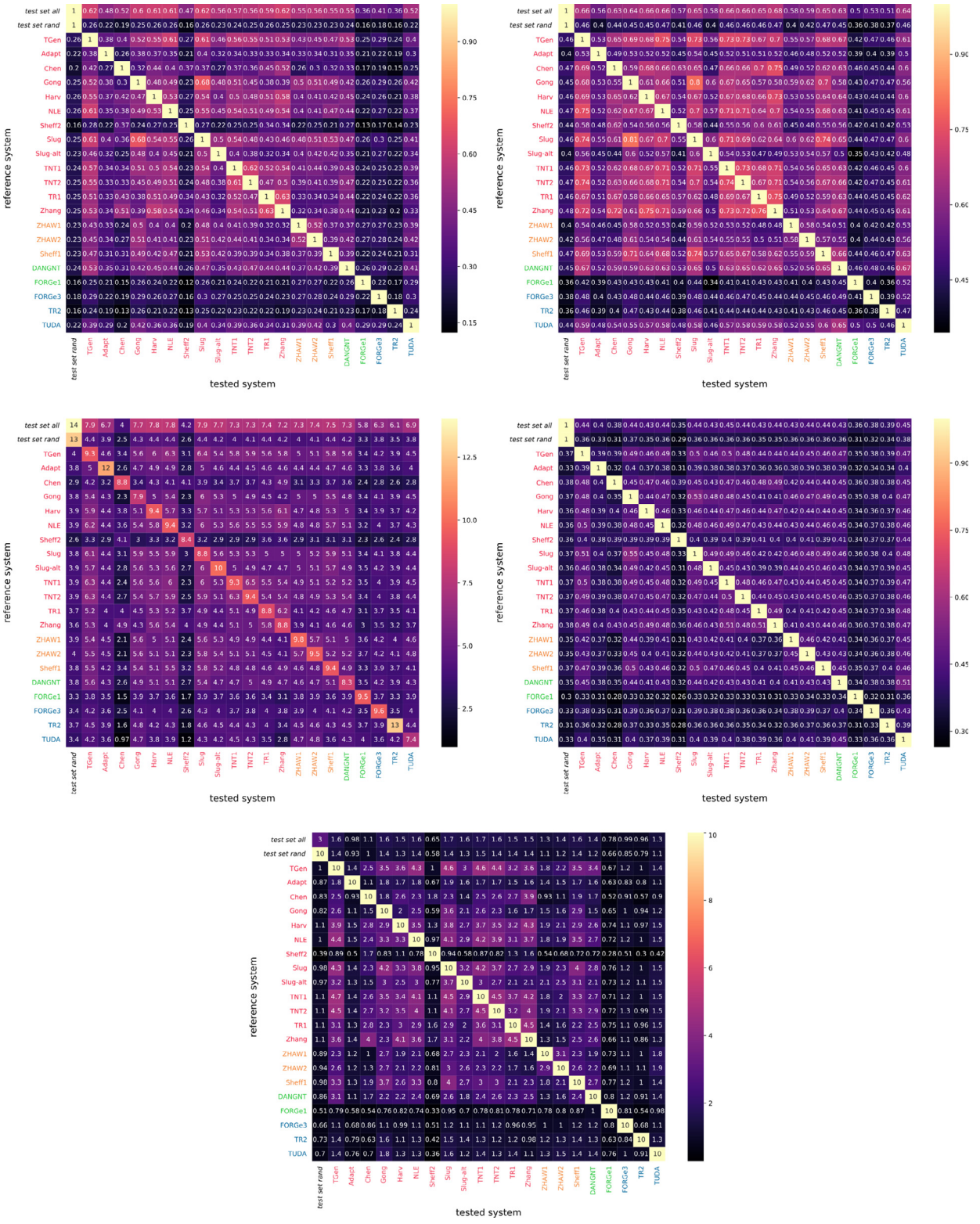


Fig. A.6. Similarity of the systems' outputs as measured by automatic metrics (left-to-right, top-to-bottom: BLEU, NIST, METEOR, ROUGE-L and CIDEr), where one of the systems is used as a reference. Systems within the graphs are sorted by their architecture. For comparison, we also include metrics values against the full test set with multiple human references and against a single (randomly chosen) test set human reference.

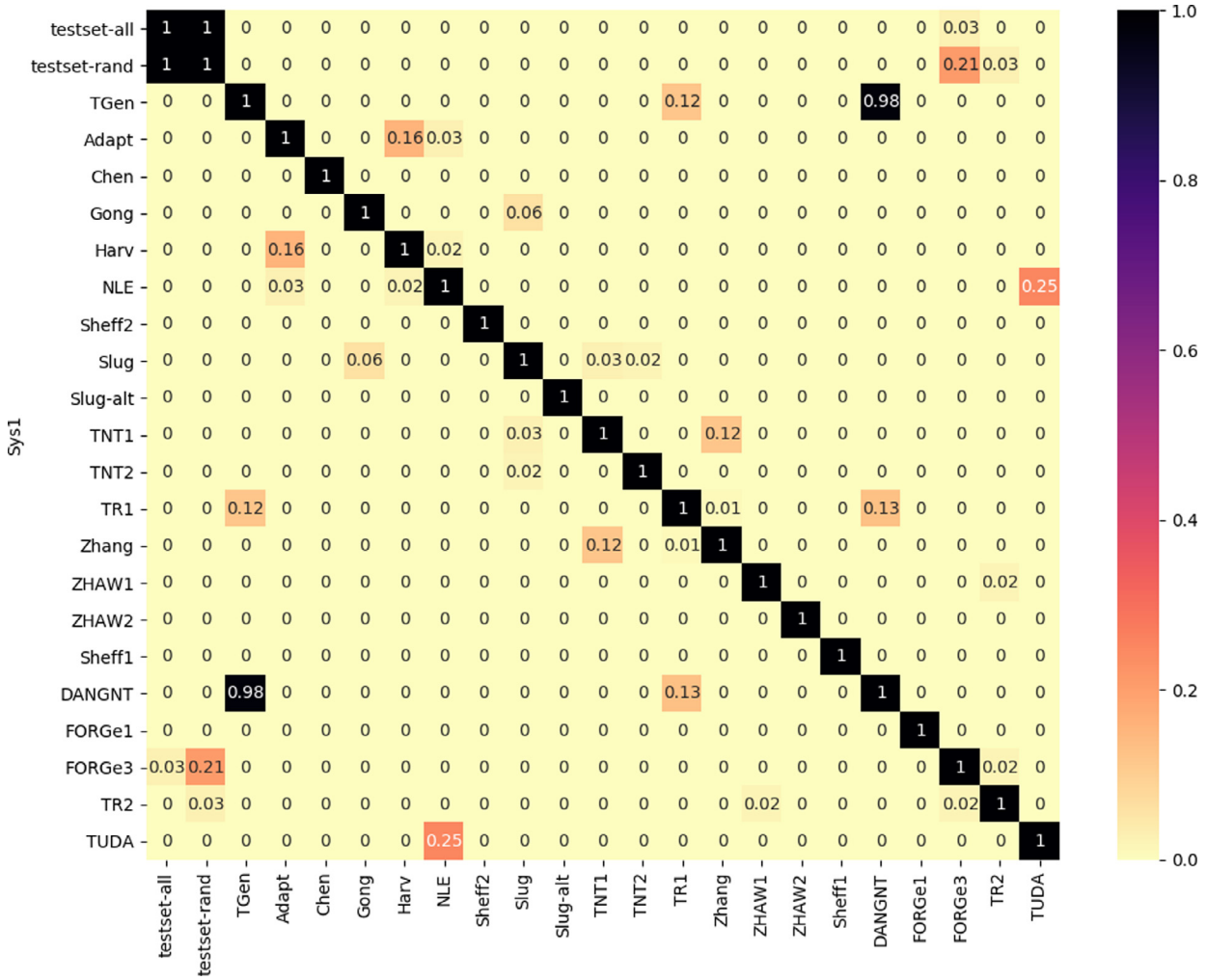


Fig. A.7. *p*-Values of the Kolmogorov–Smirnov test for discrete distributions (Arnold and Emerson, 2011), evaluating significance of differences between systems in terms of syntactic complexity of their output. Bright colour indicates statistically significant difference ($p < 0.05$).

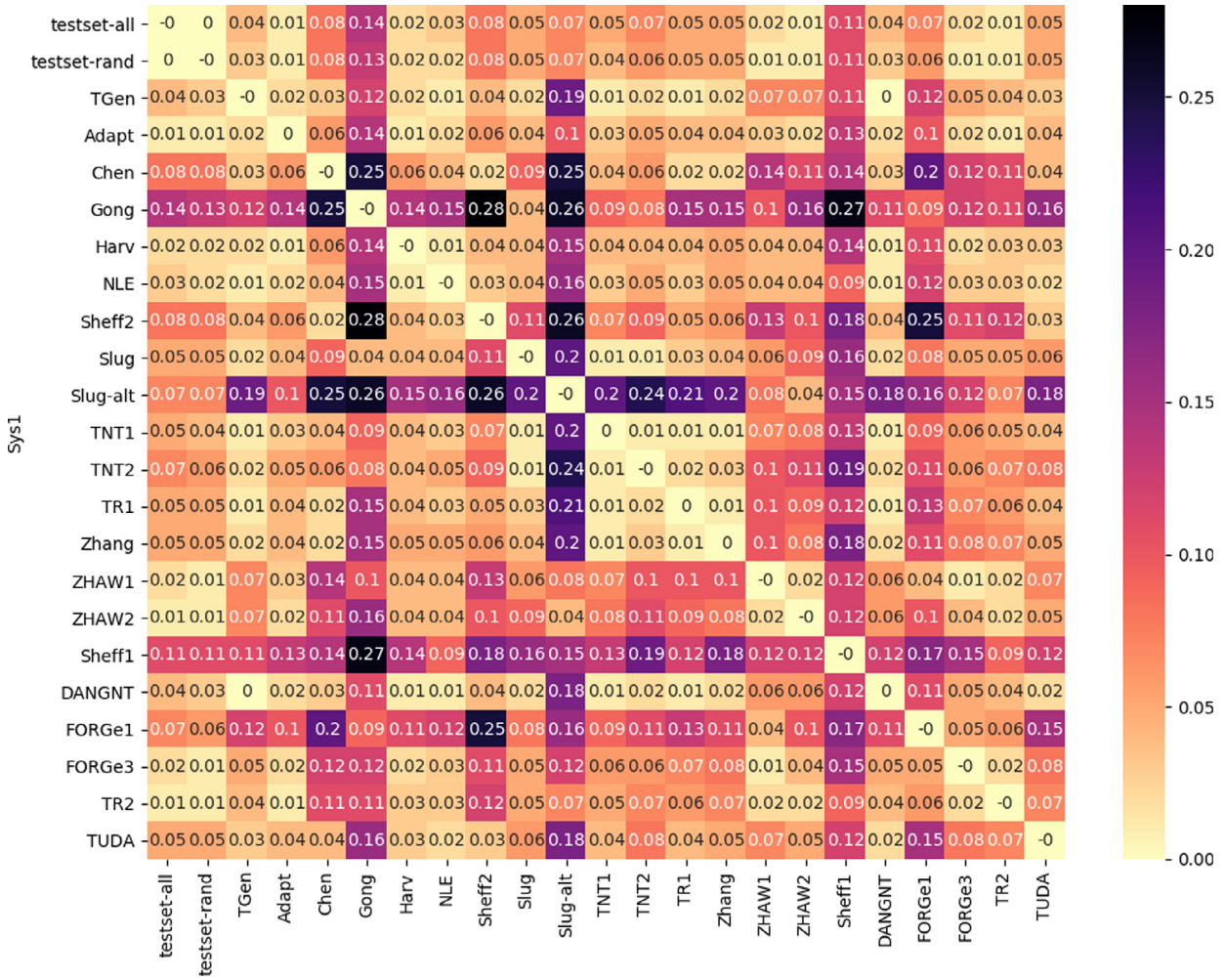


Fig. A.8. Similarities between systems, calculated using Bhattacharyya distance. Darker colour indicates greater distance, i.e. more different systems.

Table A.15
Pearson correlation between automatic metrics and human scores of naturalness. "*" denotes statistical significance at $p < 0.05$ level, bold denotes the highest value.

System name	BLEU	NIST	METEOR	ROUGE-L	CIDEr
♥TGen	0.08	0.00	0.03	0.04	0.03
♥Adapt	0.05	0.05	0.05	0.09	0.09
♥Chen	0.03	-0.03	-0.06	0.01	-0.08
♣DANGNT	0.06	-0.11	0.02	0.06	0.08
♣FORGe1	-0.06	-0.13*	-0.03	0.06	0.05
♣FORGe3	0.03	0.02	-0.03	-0.02	0.04
♥Gong	0.08	0.10	0.00	0.08	-0.02
♥Harv	0.04	0.05	0.02	0.06	-0.09
♥NLE	0.07	0.08	0.10	0.05	0.11
♦Sheff1	0.07	0.11	0.01	0.03	-0.12*
♥Sheff2	-0.11	-0.08	-0.08	-0.04	0.02
♥Slug	0.02	0.08	-0.07	0.03	-0.05
♥Slug-alt	-0.02	-0.03	0.05	0.02	0.01
♥TR1	0.15*	0.13*	0.15*	0.15*	0.02
♣TR2	0.02	0.00	0.08	0.07	0.05
♥TNT1	-0.07	-0.01	-0.08	-0.02	-0.08
♥TNT2	0.04	0.07	0.02	0.03	-0.02
♣TUDA	-0.02	-0.03	0.13	-0.04	-0.01
♥Zhang	0.03	0.01	0.03	0.00	-0.04
♦ZHAW1	0.05	0.00	0.05	0.08	0.01
♦ZHAW2	**0.16***	0.12*	0.09	0.10	0.02

Table A.16
Pearson correlation between automatic metrics and human scores of quality. "*" denotes statistical significance at $p < 0.05$ level, bold denotes the highest value.

System name	BLEU	NIST	METEOR	ROUGE-L	CIDEr
♥TGen	-0.08	-0.10	-0.05	-0.02	-0.05
♥Adapt	0.11*	0.09	0.07	0.10	0.10
♥Chen	0.07	**0.19***	0.00	0.04	0.08
♣DANGNT	-0.06	0.00	-0.08	-0.07	-0.13*
♣FORGe1	-0.01	-0.02	0.06	-0.01	0.08
♣FORGe3	0.00	-0.01	0.07	0.14*	0.04
♥Gong	0.01	-0.03	-0.03	0.02	-0.01
♥Harv	0.01	0.15*	0.05	-0.01	0.16*
♥NLE	0.09	0.03	0.05	0.09	0.15*
♦Sheff1	0.08	0.02	0.06	0.11	0.05
♥Sheff2	0.07	0.17*	0.10	0.00	0.16*
♥Slug	0.04	0.00	0.03	-0.01	0.06
♥Slug-alt	0.07	0.01	0.01	0.02	0.08
♥TR1	0.07	0.16*	-0.04	0.02	0.08
♣TR2	0.04	-0.02	0.09	0.05	0.08
♥TNT1	0.04	-0.04	-0.01	0.00	0.02
♥TNT2	0.05	0.06	0.06	0.05	0.08
♣TUDA	0.01	0.01	0.02	0.05	0.01
♥Zhang	0.02	0.16*	0.05	0.01	0.10
♦ZHAW1	-0.11	-0.05	0.00	-0.10	-0.10
♦ZHAW2	0.05	0.02	0.01	0.09	0.05

Supplementary material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.1016/j.csl.2019.06.009.

References

Agarwal, S., Dymetman, M., Gaussier, E., 2018. Char2char generation with reranking for the E2E NLG challenge. In: Proceedings of the Eleventh International Conference on Natural Language Generation. Tilburg, The Netherlands, 451–456, https://www.aclweb.org/anthology/W18-6555.

Arnold, T.B., Emerson, J.W., 2011. Nonparametric goodness-of-fit tests for discrete null distributions. R J. 3 (2), 34–39.

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: Proceedings of the Third International Conference on Learning Representations (ICLR). San Diego, CA, USA. arXiv:1409.0473.

Balakrishnan, A., Rao, J., Upasani, K., White, M., Subba, R., 2019. Constrained decoding for neural NLG from compositional representations in task-oriented dialogue. In: Proceedings of the Fifty-Seventh Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. arXiv:1906.07220.

Bao, J., Tang, D., Duan, N., Yan, Z., Lv, Y., Zhou, M., Zhao, T., 2018. Table-to-text: describing table region with natural language. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, LA, USA, pp. 5020–5027. arXiv:1805.11234.

Bard, E.G., Robertson, D., Sorace, A., 1996. Magnitude estimation of linguistic acceptability. Language 72, 32–68. doi: 10.2307/416793.

Belz, A., Gatt, A., 2007. The attribute selection for GRE challenge: overview and evaluation results. In: Proceedings of the Machine Translation Summit XI, pp. 75–83.

Belz, A., Hastie, H., 2014. Comparative evaluation and shared tasks for NLG in interactive systems. In: Stent, A., Bangalore, S. (Eds.), Natural Language Generation in Interactive Systems. Cambridge University Press, Cambridge, (Chapter 13), pp. 302–350.

Belz, A., Kow, E., 2011. Discrete vs. continuous rating scales for language evaluation in NLP. In: Proceedings of the Forty-Ninth Annual Meeting of the Association for Computational Linguistics: Short Papers. Portland, OR, USA, pp. 230–235.

Black, A.W., Burger, S., Conkie, A., Hastie, H., Keizer, S., Lemon, O., Merigaud, N., Parent, G., Schubiner, G., Thomson, B., Williams, J.D., Yu, K., Young, S., Eskenazi, M., 2011. Spoken dialog challenge 2010: comparison of live and control test results. In: Proceedings of the SIGDIAL 2011 Conference. Association for Computational Linguistics, Portland, Oregon, pp. 2–7.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., Specia, L., 2013. Findings of the 2013 workshop on statistical machine translation. In: Proceedings of the Eighth Workshop on Statistical Machine Translation. Association for Computational Linguistics, Sofia, Bulgaria, pp. 1–44.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., Tamchyna, A., 2014. Findings of the 2014 workshop on statistical machine translation. In: Proceedings of the Ninth Workshop on Statisticals Machine Translation. Association for Computational Linguistics, Baltimore, Maryland, USA, pp. 12–58.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., et al., 2017a. Findings of the 2017 conference on machine translation (WMT17). In: Proceedings of the Second Conference on Machine Translation (WMT). Copenhagen, Denmark, pp. 169–214.

Bojar, O., Graham, Y., Kamran, A., 2017b. Results of the WMT17 metrics shared task. In: Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers. Association for Computational Linguistics, Copenhagen, Denmark, pp. 489–513.

Britz, D., Goldie, A., Luong, M.-T., Le, Q., 2017. Massive exploration of neural machine translation architectures. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, pp. 1442–1451. arXiv:1703.03906.

Callison-Burch, C., Dredze, M., 2010. Creating speech and language data with Amazon's mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. Association for Computational Linguistics, pp. 1–12.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J., 2007. (Meta-) evaluation of machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation (WMT). Prague, Czech Republic, pp. 136–158.

Chang, K.-W., Krishnamurthy, A., Agarwal, A., Daumé III, H.D., Langford, J., 2015. Learning to search better than your teacher. In: Proceedings of the International Conference on Machine Learning (ICML). Lille, France. arXiv:1502.02206.

Chen, D.L., Mooney, R.J., 2008. Learning to sportscast: a test of grounded language acquisition. In: Proceedings of the Twenty-Fifth International Conference on Machine Learning (ICML). Helsinki, Finland, pp. 128–135.

Chen, M., Lampouras, G., Vlachos, A., 2018. Sheffield at E2E: structured prediction approaches to end-to-end language generation. In: Proceedings of the E2E NLG Challenge System Descriptions.

Chen, S., 2018. A general model for neural text generation from structured data. In: Proceedings of the E2E NLG Challenge System Descriptions.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L., 2015. Microsoft COCO captions: data collection and evaluation server. CoRR. arXiv:1504.00325.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, pp. 1724–1734. arXiv:1406.1078.

Collins, M., 1997. Three generative, lexicalised models for statistical parsing. In: Proceedings of the Eighth Conference of the European Chapter of the Association for Computational Linguistics (EACL). Madrid, Spain, pp. 16–23.

Covington, M.A., He, C., Brown, C., Naçi, L., Brown, J., 2006. How Complex Is That Sentence? A Proposed Revision of the Rosenberg and Abbeduto D-Level Scale. Technical Report CASPR Research Report 2006-01. University of Georgia, Athens, GA, USA.

Crammer, K., Kulesza, A., Dredze, M., 2009. Adaptive regularization of weight vectors. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, pp. 414–422.

Deriu, J., Cieliebak, M., 2018a. End-to-end trainable system for enhancing diversity in natural language generation. In: Proceedings of the E2E NLG Challenge System Descriptions.

Deriu, J.M., Cieliebak, M., 2018b. Syntactic manipulation for generating more diverse and interesting texts. In: Proceedings of the Eleventh International Conference on Natural Language Generation. Association for Computational Linguistics, Tilburg University, The Netherlands, pp. 22–34.

Dethlefs, N., Hastie, H., Rieser, V., Lemon, O., 2012. Optimising incremental dialogue decisions using information density for interactive systems. In: Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 82–93.

Doddington, G., 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: Proceedings of the Second International Conference on Human Language Technology Research. San Diego, CA, USA, pp. 138–145.

Dušek, O., Jurčíček, F., 2015. Training a natural language generator from unaligned data. In: Proceedings of the Fifty-Third Annual Meeting of the Association for Computational Linguistics and the Seventh International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China, pp. 451–461.

Dušek, O., 2017. Novel Methods for Natural Language Generation in Spoken Dialogue Systems. Ph.D. thesis. Charles University, Prague, Czech Republic.

Dušek, O., Jurčíček, F., 2016. A context-aware natural language generator for dialogue systems. In: Proceedings of the Seventeenth Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics, Los Angeles, CA, USA, pp. 185–190.

Dušek, O., Jurčíček, F., 2016a. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In: Proceedings of the Fifty-Fourth Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, pp. 45–51. arXiv:1606.05491.

Dušek, O., Novikova, J., Rieser, V., 2017. Referenceless quality estimation for natural language generation. In: Proceedings of the First Workshop on Learning to Generate Natural Language (LGNL). Sydney, Australia. arXiv:1708.01759.

Dušek, O., Novikova, J., Rieser, V., 2018. Findings of the E2E NLG challenge. In: Proceedings of the Eleventh International Conference on Natural Language Generation. Tilburg, The Netherlands, pp. 322–328.

Elder, H., Gehrmann, S., O'Connor, A., Liu, Q., 2018. E2E NLG challenge submission: towards controllable generation of diverse natural language. In: Proceedings of the Eleventh International Conference on Natural Language Generation. Tilburg, The Netherlands, pp. 457–462.
Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P., Boyd-Graber, J., 2018. Pathologies of neural models make interpretations difficult. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, pp. 3719–3728. doi: 10.18653/v1/D18-1407.
Freitag, M., Roy, S., 2018. Unsupervised natural language generation with denoising autoencoders. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, pp. 3922–3929. arXiv:1804.07899.
Gardent, C., Shimorina, A., Narayan, S., Perez-Beltrachini, L., 2017. The WebNLG challenge: generating text from RDF data. In: Proceedings of the 10th International Conference on Natural Language Generation. Association for Computational Linguistics, Santiago de Compostela, Spain, pp. 124–133.
Gatt, A., Krahmer, E., 2018. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. J. Artif. Intell. Res. (JAIR) 61, 65–170.
Gehrmann, S., Dai, F.Z., Elder, H., Rush, A.M., 2018. End-to-end content and plan selection for natural language generation. In: Proceedings of the Eleventh International Conference on Natural Language Generation. Tilburg, The Netherlands, pp. 46–56.
Gkatzia, D., Lemon, O., Rieser, V., 2016. Natural language generation enhances human decision-making with uncertain information. In: Proceedings of the Fifty-Fourth Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Berlin, Germany, pp. 264–268. arXiv:1606.03254.
Gkatzia, D., Mahamood, S., 2015. A snapshot of NLG evaluation practices 2005−2014. In: Proceedings of the Fifteenth European Workshop on Natural Language Generation (ENLG). Association for Computational Linguistics, Brighton, UK, pp. 57–60.
Gong, H., 2018. Technical report for E2E NLG challenge. In: Proceedings of the E2E NLG Challenge System Descriptions.
Graham, Y., Baldwin, T., Moffat, A., Zobel, J., 2013. Continuous measurement scales in human evaluation of machine translation. In: Proceedings of the Seventh Linguistic Annotation Workshop & Interoperability with Discourse. Sofia, Bulgaria, pp. 33–41.
Guzman-Rivera, A., Batra, D., Kohli, P., 2012. Multiple choice learning: learning to produce multiple structured outputs. In: Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, NV, USA, pp. 1799–1807.
Han, L., Kashyap, A., Finin, T., Mayfield, J., Weese, J., 2013. UMBC_EBIQUITY-CORE: semantic textual similarity systems. In: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), 1. Atlanta, Georgia, pp. 44–52.
Henderson, M., Thomson, B., Williams, J.D., 2014. The second dialog state tracking challenge. In: Proceedings of the Fifteenth Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). Association for Computational Linguistics, Philadelphia, PA, U.S.A., pp. 263–272.
Herbrich, R., Minka, T., Graepel, T., 2006. Trueskill™: a Bayesian skill rating system. In: Proceedings of the Nineteenth International Conference on Neural Iinformation Processing Systems (NIPS). Vancouver, Canada, pp. 569–576.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Jagfeld, G., Jenne, S., Vu, N.T., 2018. Sequence-to-sequence models for data-to-text natural language generation: word- vs. character-based processing and output diversity. In: Proceedings of the Eleventh International Conference on Natural Language Generation. Tilburg, The Netherlands, pp. 221–232. arXiv:1810.04864.
Juraska, J., Karagiannis, P., Bowden, K.K., Walker., M.A., 2018. A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. In: Proceedings of the 2018 NAACL-HLT. New Orleans, LA, USA, pp. 152–162.
Juraska, J., Walker, M., 2018. Characterizing variation in crowd-sourced data for training neural language generators to produce stylistically varied outputs. In: Proceedings of the Eleventh International Conference on Natural Language Generation. Tilburg, The Netherlands, pp. 441–450. arXiv:1809.05288.
Kaffee, L.-A., Elsahar, H., Vougiouklis, P., Gravier, C., Laforest, F., Hare, J., Simperl, E., 2018. Learning to generate Wikipedia summaries for underserved languages from Wikidata. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, LA, USA, pp. 640–645. arXiv:1803.07116.
Kann, K., Rothe, S., Filippova, K., 2018. Sentence-level fluency evaluation: references help, but can be spared!. In: Proceedings of the Twenty-Second Conference on Computational Natural Language Learning. Brussels, Belgium, pp. 313–323.
Kiddon, C., Zettlemoyer, L., Choi, Y., 2016. Globally coherent text generation with neural checklist models. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA, pp. 329–339. doi: 10.18653/v1/D16-1032.
Kingma, D., Ba, J., 2015. Adam: a method for stochastic optimization. In: Proceedings of the Third International Conference on Learning Representations. San Diego, CA, USA. arXiv:1412.6980 .
Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A., 2017. OpenNMT: open-source toolkit for neural machine translation. In: Proceedings of the Fifty-Fifth Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada, pp. 67–72. doi: 10.18653/v1/P17-4012.
Koehn, P., Knowles, R., 2017. Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation. Vancouver, Canada, pp. 28–39. doi: 10.18653/v1/W17-3204.
Lampouras, G., Vlachos, A., 2016. Imitation learning for language generation from unaligned data. In: Proceedings of the Twenty-Sixth International Conference on Computational Linguistics: Technical Papers, COLING 2016. The COLING 2016 Organizing Committee, Osaka, Japan, pp. 1101–1112.
Lavie, A., Agarwal, A., 2007. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation. Prague, Czech Republic, pp. 228–231.
Lebret, R., Grangier, D., Auli, M., 2016. Neural text generation from structured data with application to the biography domain. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA, pp. 1203–1213. arXiv:1603.07771.
Lee, K., Firat, O., Agarwal, A., Fannjiang, C., Sussillo, D., 2019. Hallucinations in Neural Machine Translation. https://openreview.net/forum?id=SkxJ-309FQ.
Lin, C.-Y., 2004. ROUGE: a package for automatic evaluation of summaries. Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. Barcelona, Spain, pp. 74–81.
Liu, B., Lane, I., 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. In: Proceedings of the 2016 INTERSPEECH. San Francisco, CA, USA, pp. 685–689. arXiv:1609.01454.
Lu, X., 2009. Automatic measurement of syntactic complexity in child language acquisition. Int. J. Corpus Linguist. 14 (1), 3–28.
Lu, X., 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. Mod. Lang. J. 96 (2), 190–208.
Mairesse, F., Gašić, M., Jurčíček, F., Keizer, S., Thomson, B., Yu, K., Young, S., 2010. Phrase-based statistical language generation using graphical models and active learning. In: Proceedings of the Forty-Eighth Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, pp. 1552–1561.
Mairesse, F., Walker, M., 2007. PERSONAGE: personality generation for dialogue. In: Proceedings of the Forty-Fifth Annual Meeting of the Association For Computational Linguistics. Prague, pp. 496–503.
Mairesse, F., Walker, M.A., 2011. Controlling user perceptions of linguistic style: trainable generation of personality traits. Comput. Linguist. 37 (3), 455–488. doi: 10.1162/COLI_a_00063.
Mangrulkar, S., Shrivastava, S., Thenkanidiyoor, V., Dinesh, D.A., 2018. A context-aware convolutional natural language generation model for dialogue systems. In: Proceedings of the Nineteenth Annual SIGdial Meeting on Discourse and Dialogue. Melbourne, Australia, pp. 191–200.
Manishina, E., Jabaian, B., Huet, S., Lefevre, F., 2016. Automatic corpus extension for data-driven natural language generation. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC). Portorož, Slovenia, pp. 3624–3631.
Manning, C.D., Schütze, H., 2000. Foundations of Statistical Natural Language Processing, 2nd printing MIT Press, Cambridge, MA, USA.
Mason, W., Watts, D.J., 2010. Financial incentives and the performance of crowds. ACM SigKDD Explor. Newslett. 11 (2), 100–108.
Mei, H., Bansal, M., Walter, M.R., 2016. What to talk about and how? Selective generation using LSTMs with coarse-to-fine alignment. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, CA, USA, pp. 720–730. arXiv:1509.00838 .
Mel'čuk, I.A., 1988. Dependency Syntax: Theory and Practice. SUNY Series in Linguistics. State University Press of New York, Albany, NY, USA.
Mille, S., Dasiopoulou, S., 2018. FORGe at E2E 2017. In: Proceedings of the E2E NLG Challenge System Descriptions.
Nayak, N., Hakkani-Tür, D., Walker, M., Heck, L., 2017. To plan or not to plan? Discourse planning in slot-value informed sequence to sequence models for language generation. In: Proceedings of the 2017 INTERSPEECH. Stockholm, Sweden, pp. 3339–3343. doi: 10.21437/Interspeech.2017-1525.

Nema, P., Shetty, S., Jain, P., Laha, A., Sankaranarayanan, K., Khapra, M.M., 2018. Generating descriptions from structured data using a bifocal attention mechanism and gated orthogonalization. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, LA, USA, pp. 1539–1550. doi: 10.18653/v1/N18-1139. arXiv:1804.07789.
Nguyen, D.T., Tran, T., 2018. Structure-based generation system for E2E NLG challenge. In: Proceedings of the E2E NLG Challenge System Descriptions.
Novikova, J., Dušek, O., Cercas Curry, A., Rieser, V., 2017a. Why we need new evaluation metrics for NLG. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2231–2242.
Novikova, J., Dušek, O., Rieser, V., 2017b. The E2E dataset: new challenges for end-to-end generation. In: Proceedings of the Eighteenth Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL), pp. 201–206.
Novikova, J., Dušek, O., Rieser, V., 2018. RankME: reliable human ratings for natural language generation. In: Proceedings of the Sixteenth Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). New Orleans, LA, USA, pp. 72–78. arXiv:1803.05928.
Novikova, J., Lemon, O., Rieser, V., 2016. Crowd-sourcing NLG data: pictures elicit better data. In: Proceedings of the Ninth International Natural Language Generation Conference. Edinburgh, UK, pp. 265–273. arXiv:1608.00339.
Novikova, J., Rieser, V., 2016. The aNALoGuE Challenge: Non Aligned Language GEneration. In: Proceedings of the Ninth International Natural Language Generation Conference, pp. 168–170.
Oraby, S., Harrison, V., Ebrahimi, A., Walker, M., 2019. Curate and generate: a corpus and method for joint control of semantics and style in neural NLG. In: Proceedings of the Fifty-Seventh Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. arXiv:1906.01334.
Oraby, S., Reed, L., Tandon, S., 2018a. Controlling personality-based stylistic variation with neural natural language generators. In: Proceedings of the Nineteenth Annual SIGdial Meeting on Discourse and Dialogue. Melbourne, Australia, pp. 180–190.
Oraby, S., Reed, L., Tandon, S., Sharath, T.S., Lukin, S., Walker, M., 2018b. TNT-NLG, system 1: using a statistical NLG to massively augment crowd-sourced data for neural generation. In: Proceedings of the E2E NLG Challenge System Descriptions.
Oraby, S., Reed, L., Sharath, T.S., Tandon, S., Walker, M., 2018c. Neural multivoice models for expressing novel personalities in dialog. In: Proceedings of the 2018 INTERSPEECH. Hyderabad, India, pp. 3057–3061. doi: 10.21437/Interspeech.2018-2174. arXiv:1809.01331.
Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the Fortieth Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, PA, USA, pp. 311–318.
Parra Escartín, C., Reijers, W., Lynn, T., Moorkens, J., Way, A., Liu, C.-H., 2017. Ethical considerations in NLP shared tasks. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. Association for Computational Linguistics, pp. 66–73.
Perez-Beltrachini, L., Gardent, C., 2017. Analysing data-to-text generation benchmarks. In: Proceedings of the Tenth International Natural Language Generation Conference. Santiago de Compostela, Spain, pp. 238–242.
Puzikov, Y., Gurevych, I., 2018. E2E NLG challenge: neural models vs. templates. In: Proceedings of the Eleventh International Conference on Natural Language Generation. Tilburg, The Netherlands, pp. 463–471.
Reed, L., Oraby, S., Walker, M., 2018. Can neural generators for dialogue learn sentence planning and discourse structuring? In: Proceedings of the Eleventh International Conference on Natural Language Generation. Tilburg, The Netherlands, pp. 284–295. arXiv:1809.03015.
Reiter, E., 2018. A structured review of the validity of BLEU. Comput. Linguist. 44 (3), 393–401. doi: 10.1162/COLI_a_00322.
Reiter, E., 2019. Does Deep Learning Prefer Readability over Accuracy? Ehud Reiter's Blog. Available online at https://ehudreiter.com/2019/01/08/deep-learning-prefer-readability/ (accessed: Jan 10, 2019)
Reiter, E., Belz, A., 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Comput. Linguist. 35 (4), 529–558. doi: 10.1162/coli.2009.35.4.35405.
Rieser, V., Lemon, O., 2009. Natural language generation as planning under uncertainty for spoken dialogue systems. In: Proceedings of the Twelfth Conference of the European Chapter of the ACL (EACL). Athens, Greece, pp. 683–691.
Rieser, V., Lemon, O., Keizer, S., 2014. Natural language generation as incremental planning under uncertainty: adaptive information presentation for statistical dialogue systems. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (5), 979–993. doi: 10.1109/TASL.2014.2315271.
Roberti, M., Bonetta, G., Cancelliere, R., Gallinari, P., 2019. Copy mechanism and tailored training for character-based data-to-text generation. In: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Würzburg, Germany. arXiv:1904.11838.
Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K., 2018. Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, pp. 4035–4045.
Sakaguchi, K., Post, M., Van Durme, B., 2014. Efficient elicitation of annotations for human evaluation of machine translation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT). Baltimore, MD, USA, pp. 1–11.
Sennrich, R., Haddow, B., Birch, A., 2016. Neural machine translation of rare words with subword units. In: Proceedings of the Fifty-Fourth Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany, pp. 1715–1725. arXiv:1508.07909.
Sharma, S., He, J., Suleman, K., Schulz, H., Bachman, P., 2016. Natural language generation in dialogue using lexicalized and delexicalized data. CoRR. arXiv:1606.03632.
Shimorina, A., Gardent, C., 2018. Handling rare items in data-to-text generation. In: Proceedings of the Eleventh International Conference on Natural Language Generation. Tilburg, The Netherlands, pp. 360–370.
Smiley, C., Davoodi, E., Song, D., Schilder, F., 2018. The E2E NLG challenge: a tale of two systems. In: Proceedings of the Eleventh International Conference on Natural Language Generation. Tilburg, The Netherlands, pp. 472–477.
Specia, L., Raj, D., Turchi, M., 2010. Machine translation evaluation versus quality estimation. Machine Transl. 24 (1), 39–50.
Sprouse, J., 2011. A validation of Amazon mechanical turk for the collection of acceptability judgments in linguistic theory. Behav. Res. Methods 43 (1), 155–167.
Sripada, S.G., Reiter, E., Hawizy, L., 2005. Evaluation of an NLG system using post-edit data: lessons learnt. In: Proceedings of the Tenth European Workshop on Natural Language Generation.
Stent, A., Prasad, R., Walker, M., 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In: Proceedings of the Forty-Second Meeting of the Association for Computational Linguistics. Barcelona, Spain, pp. 79–86.
Straková, J., Straka, M., Hajič, J., 2014. Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In: Proceedings of Fifty-Second Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Association for Computational Linguistics, Baltimore, Maryland, pp. 13–18.
Su, S.-Y., Chen, Y.-N., 2018. Investigating linguistic pattern ordering in hierarchical natural language generation. In: Proceedings of the IEEE Spoken Language Technology Workshop. Athens, Greece. arXiv:1809.07629.
Su, S.-Y., Huang, C.-W., Chen, Y.-N., 2019. Dual supervised learning for natural language understanding and generation. In: Proceedings of the Fifty-Seventh Annual Meeting of the Association for Computational Linguistics (ACL). Florence, Italy. arXiv:1905.06196.
Su, S.-Y., Lo, K.-L., Yeh, Y.-T., Chen, Y.-N., 2018. Natural language generation by hierarchical decoding with linguistic patterns. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, LA, USA, pp. 61–66.
Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 3104–3112. arXiv:1409.3215.
Tandon, S., Sharath, T.S., Oraby, S., Reed, L., Lukin, S., Walker, M., 2018. TNT-NLG, system 2: data repetition and meaning representation manipulation to improve neural generation. In: Proceedings of the E2E NLG Challenge System Descriptions.
Tian, Y., Douratsos, I., Groves, I., 2018. Treat the system like a human student: Automatic naturalness evaluation of generated text without reference texts. In: Proceedings of the Eleventh International Conference on Natural Language Generation. Tilburg, The Netherlands, pp. 109–118.

Tran, V.-K., Nguyen, L.-M., Tojo, S., 2017. Neural-based natural language generation in dialogue using RNN encoder–decoder with semantic aggregation. In: Proceedings of the Eighteenth Annual SIGdial Meeting on Discourse and Dialogue. Saarbrücken, Germany, pp. 231–240. doi: 10.18653/v1/W17-5528. arXiv:1706.06714.

Ueffing, N., Camargo de Souza, J.G., Leusch, G., 2018. Quality estimation for automatically generated titles of eCommerce browse pages. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers). Association for Computational Linguistics, New Orleans, LA, USA, pp. 52–59. doi: 10.18653/v1/N18-3007.

Vedantam, R., Zitnick, C.L., Parikh, D., 2015. CIDEr: consensus-based image description evaluation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA, pp. 4566–4575. doi: 10.1109/CVPR.2015.7299087.

Vinyals, O., Fortunato, M., Jaitly, N., 2015. Pointer networks. In: Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015). Montréal, Canada. arXiv:1506.03134.

Walker, M.A., Whittaker, S.J., Stent, A., Maloor, P., Moore, J., Johnston, M., Vasireddy, G., 2004. Generation and evaluation of user tailored responses in multimodal dialogue. Cognit. Sci. 28 (5), 811–840.

Wang, Q., Pan, X., Huang, L., Zhang, B., Jiang, Z., Ji, H., Knight, K., 2018. Describing a knowledge base. In: Proceedings of the Eleventh International Conference on Natural Language Generation. Tilburg University, The Netherlands, pp. 10–21.

Wang, W.Y., Bohus, D., Kamar, E., Horvitz, E., 2012. Crowdsourcing the acquisition of natural language corpora: Methods and observations. In: Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 73–78.

Wang, Y., Berant, J., Liang, P., 2015. Building a semantic parser overnight. In: Proceedings of the Fifty-Third Annual Meeting of the Association for Computational Linguistics and the Seventh International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Beijing, China, pp. 1332–1342.

Wen, T., Gašić, M., Mrkšić, N., Rojas-Barahona, L.M., Su, P., Vandyke, D., Young, S.J., 2016. Multi-domain neural network language generation for spoken dialogue systems. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, CA, USA, pp. 120–129. arXiv:1603.01232.

Wen, T.-H., Gasić, M., Kim, D., Mrkšić, N., Su, P.-H., Vandyke, D., Young, S., 2015. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In: Proceedings of the Sixteenth Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). Association for Computational Linguistics, Prague, Czech Republic, pp. 275–284.

Wen, T.-H., Gašić, M., Mrkšić, N., Su, P.-H., Vandyke, D., Young, S., 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, pp. 1711–1721.

Wen, T.-H., Vandyke, D., Mrkšić, N., Gašić, M., Rojas-Barahona, L.M., Su, P.-H., Ultes, S., Young, S., 2017. A network-based end-to-end trainable task-oriented dialogue system. In: Proceedings of the Fifteenth Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Valencia, Spain, pp. 438–449. arXiv:1604.04562.

Williams, J.D., Young, S., 2007. Partially observable Markov decision processes for spoken dialog systems. Comput. Speech Lang. 21 (2), 393–422. doi: 10.1016/j.csl.2006.06.008.

Williams, R.J., 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. 8, 229–256.

Wiseman, S., Shieber, S.M., Rush, A.M., 2017. Challenges in data-to-document generation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017. Copenhagen, Denmark, September 9–11, 2017, pp. 2253–2263.

Wiseman, S., Shieber, S.M., Rush, A.M., 2018. Learning neural templates for text generation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, pp. 3174–3187. arXiv:1808.10122.

Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., Yu, K., 2010. The hidden Information State model: a practical framework for POMDP-based spoken dialogue management. Comput. Speech Lang. 24 (2), 150–174. doi: 10.1016/j.csl.2009.04.001.

Zaidan, O.F., Callison-Burch, C., 2011. Crowdsourcing translation: professional quality from non-professionals. In: Proceedings of the ACL. Portland, Oregon, USA, pp. 1220–1229.

Zhang, B., Xiong, D., Su, J., Duan, H., 2017. A context-aware recurrent encoder for neural machine translation. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (12), 2424–2432. doi: 10.1109/TASLP.2017.2751420.

Zhang, B., Yang, J., Lin, Q., Su, J., 2018. Attention regularized sequence-to-sequence learning for E2E NLG challenge. In: Proceedings of the E2E NLG Challenge System Descriptions.