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

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

Abstract

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

1 Introduction

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

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

Table 1: An example of a data instance.

In this paper, we describe a new crowdsourced dataset of 50k instances in the restaurant domain (see Section 2). We analyse it following the methodology proposed by Perez-Beltrachini and Gardent (2017) and show that the dataset brings additional challenges, such as open vocabulary, complex syntactic structures and diverse discourse phenomena, as described in Section 3. The data is openly released as part of the E2E NLG challenge.[1] We establish a baseline on the dataset in Section 4, using one of the previous end-to-end approaches.

2 The E2E Dataset

The data was collected using the CrowdFlower platform and quality-controlled following Novikova et al. (2016). The dataset provides infor-

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

Proceedings of the SIGDIAL 2017 Conference, pages 201–206,
Saarbrücken, Germany, 15-17 August 2017.

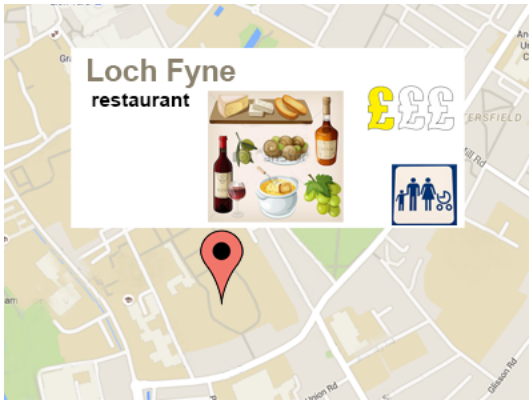

Figure 1: Pictorial MR for Table 1.

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, ...
area	dictionary	riverside, city center, ...
customerRating	enumerable	1 of 5 (low), 4 of 5 (high), ...

Table 2: Domain ontology of the E2E dataset.

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

3 Challenges

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

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

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

We also investigate the distribution of the top 25 most frequent bigrams and trigrams in our dataset (see Figure 2). The majority of both trigrams (61%) and bigrams (50%) is only used once in the dataset, which creates a challenge to efficiently train on this data. Bigrams used more than once in the dataset have an average frequency of 54.4 (SD = 433.1), and the average frequency of trigrams used more than once is 19.9 (SD = 136.9). For comparison, neither SFRest nor Bagel dataset contains bigrams or trigrams that are only used once. The minimal frequency of bigrams is 27 for Bagel (Mean = 98.2, SD = 86.9) and 76 for SFrest (Mean = 128.4, SD = 50.5), for trigrams the minimal frequency is 24 for Bagel (Mean = 63.5, SD = 54.6) and 43 for SFRest (Mean = 67.3, SD = 18.9). Infrequent words and phrases pose a chal-

[2]Note that the difference is even bigger in practice as the Refs/MR ratio for the SFRest dataset is skewed: for specific MRs, e.g. *goodbye*, SFRest has up to 101 references.

	No. of instances	No. of unique MRs	Refs/MR		Slots/MR	W/Ref	W/Sent	Sents/Ref	
E2E	**50,602**	**5,751**	**8.1**	**(2–16)**	**5.43**	**20.1**	**14.3**	**1.5**	**(1–6)**
SFRest	5,192	1,950	1.82	(1–101)	2.86	8.53	8.53	1.05	(1–4)
Bagel	404	202	2	(2–2)	5.41	11.54	11.54	1.02	(1–2)

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

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

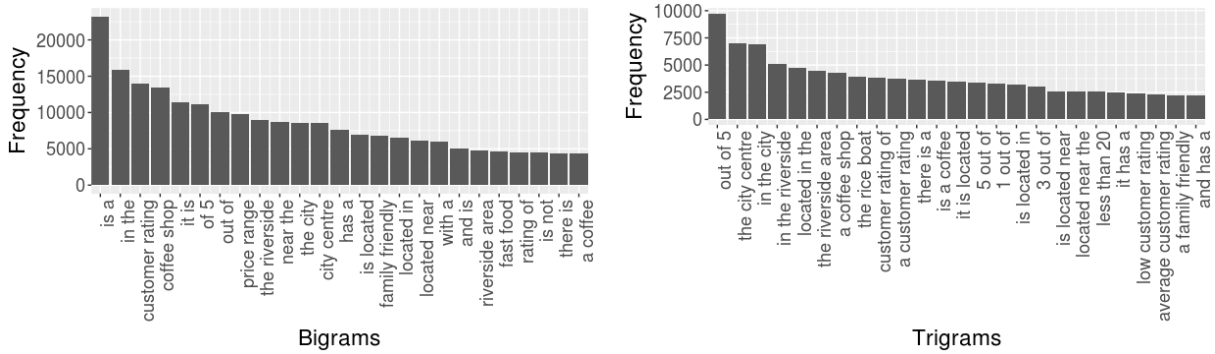

Figure 2: Distribution of the top 25 most frequent bigrams and trigrams in our dataset (left: most frequent bigrams, right: most frequent trigrams).

Dataset	Tokens	Types	LS	TTR	MSTTR
E2E	**65,710**	945	**0.57**	0.01	**0.75**
SFRest	45,791	1,187	0.43	0.03	0.62
Bagel	1,071	70	0.42	0.04	0.41

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

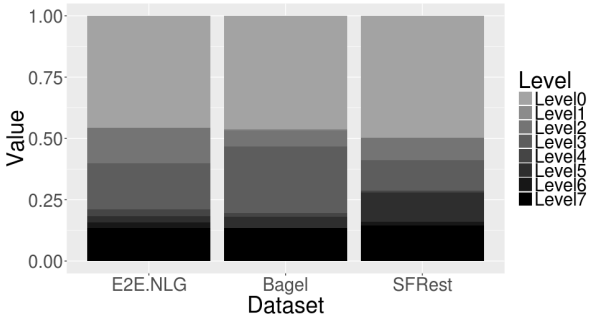

Figure 3: D-Level sentence distribution of the datasets under comparison.

lenge to current end-to-end generators since they cannot handle out-of-vocabulary words.

Syntactic Variation and Discourse Phenomena: We used the D-Level Analyser (Lu, 2009) to evaluate syntactic variation and complexity of human references using the revised D-Level Scale (Lu, 2014). Figure 3 show a similar syntactic variation in all three datasets. Most references in all the datasets are simple sentences (levels 0 and 1), although the proportion of simple texts is the lowest for the E2E NLG dataset (46%) compared to others (47-51%). Examples of simple sentences in our dataset include: "The Vaults is an Indian restaurant", or "The Loch Fyne is a moderate priced family restaurant". The majority of our data, however, contains more complex, varied syntactic structures, including phenomena explicitly modelled by early statistical approaches (Stent et al., 2004; Walker et al., 2004). For example, clauses may be joined by a coordinating conjunction (level 2), e.g. "Cocum is a very expensive restaurant *but* the quality is great". There are 14% of level-2 sentences in our dataset, comparing to 7-9% in others. Sentences may also contain verbal gerund (*-ing*) phrases (level 4), either in addition to previously discussed structures or separately, e.g. "The coffee shop Wildwood has fairly priced food, *while being* in the same vicinity as the Ranch" or "The Vaults is a family-friendly restaurant *offering* fast food at moderate prices". Subordinate clauses are marked as level 5, e.g. "*If* you like Japanese food, try the Vaults". The highest levels of syntactic complexity involve

Dataset	O	A	C
E2E NLG	22%	18%	60%
SFRest	0%	6%	94%
Bagel	0%	0%	100%

Table 5: Match between MRs and NL references.

O: Omitted content, A: Additional content, C: Content fully covered in the reference.

Metric	Value
BLEU (Papineni et al., 2002)	0.6925
NIST (Doddington, 2002)	8.4781
METEOR (Lavie and Agarwal, 2007)	0.4703
ROUGE-L (Lin, 2004)	0.7257
CIDEr (Vedantam et al., 2015)	2.3987

Table 6: TGen results on the development set.

sentences containing referring expressions ("The Golden Curry provides Chinese food in the high price range. *It* is near the Bakers"), non-finite clauses in adjunct position ("*Serving* cheap English food, as well as *having* a coffee shop, the Golden Palace has an average customer rating and is located along the riverside") or sentences with multiple structures from previous levels. All the datasets contain 13-16% of sentences of levels 6 and 7, where Bagel has the lowest proportion (13%) and our dataset the highest (16%).

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

4 Baseline System Performance

To establish a baseline on the task data, we use TGen (Dušek and Jurčíček, 2016a), one of the recent E2E data-driven systems.[3] TGen is based on sequence-to-sequence modelling with attention (seq2seq) (Bahdanau et al., 2015). In addition to the standard seq2seq model, TGen uses beam search for decoding and a reranker over the top k outputs, penalizing those outputs that do not verbalize all attributes from the input MR. As TGen does not handle unknown vocabulary well, the sparsely occurring string attributes (see Table 2) *name* and *near* are delexicalized – replaced with placeholders during generation time (both in input MRs and training sentences).[4]

We evaluated TGen on the development part of the E2E set using several automatic metrics. The results are shown in Table 6.[5] Despite the greater variety of our dataset as shown in Section 3, the BLEU score achieved by TGen is in the same range as scores reached by the same system for BAGEL (0.6276) and SFRest (0.7270). This indicates that the size of our dataset and the increased number of human references per MR helps statistical approaches.

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

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

[4] Detailed system training parameters are given in the supplementary material.

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

5 Conclusion

We described the E2E dataset for end-to-end, statistical natural language generation systems. While this dataset is ten times bigger than similar, frequently used datasets, it also poses new challenges given its lexical richness, syntactic complexity and discourse phenomena. Moreover, generating from this set also involves content selection. In contrast to previous datasets, the E2E data is crowdsourced using pictorial stimuli, which was shown to elicit more natural, more informative and better phrased human references than textual meaning representations (Novikova et al., 2016). As such, learning from this data promises more natural and varied outputs than previous "template-like" datasets. The dataset is freely available as part of the E2E NLG Shared Task.[6]

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

Acknowledgements

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

References

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*. San Diego, CA, USA. ArXiv:1409.0473. http://arxiv.org/abs/1409.0473.

David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: A test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning (ICML)*. Helsinki, Finland, pages 128–135. http://dl.acm.org/citation.cfm?id=1390173.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *CoRR* abs/1504.00325. http://arxiv.org/abs/1504.00325.

Vera Demberg and Johanna D Moore. 2006. Information presentation in spoken dialogue systems. In *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL)*. pages 65–72. http://aclweb.org/anthology/E06-1009.

George Doddington. 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, pages 138–145. http://dl.acm.org/citation.cfm?id=1289273.

Ondřej Dušek and Filip Jurčíček. 2015. Training a natural language generator from unaligned data. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China, pages 451–461. http://aclweb.org/anthology/P15-1044.

Ondřej Dušek and Filip Jurčíček. 2016a. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, pages 45–51. arXiv:1606.05491. http://aclweb.org/anthology/P16-2008.

Ondřej Dušek and Filip Jurčíček. 2016b. A context-aware natural language generator for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Los Angeles, CA, USA, pages 185–190. arXiv:1608.07076. http://aclweb.org/anthology/W16-3622.

Gerasimos Lampouras and Andreas Vlachos. 2016. Imitation learning for language generation from unaligned data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 1101–1112. http://aclweb.org/anthology/C16-1105.

Alon Lavie and Abhaya Agarwal. 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*. Association for Computational Linguistics, Prague, Czech Republic, pages 228–231. http://aclweb.org/anthology/W07-0734.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-*

[6]The training and development parts of our dataset can be downloaded from http://www.macs.hw.ac.uk/InteractionLab/E2E/.

04 workshop. Barcelona, Spain, pages 74–81. http://aclweb.org/anthology/W04-1013.

Xiaofei Lu. 2009. Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics* 14(1):3–28. http://doi.org/10.1075/ijcl.14.1.02lu.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal* 96(2):190–208. http://doi.org/10.1111/j.1540-4781.2011.01232_1.x.

Xiaofei Lu. 2014. *Computational methods for corpus annotation and analysis*. Springer. http://doi.org/10.1007/978-94-017-8645-4.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pages 1552–1561. http://aclweb.org/anthology/P10-1157.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 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. arXiv:1509.00838. http://aclweb.org/anthology/N16-1086.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing NLG data: Pictures elicit better data. In *Proceedings of the 9th International Natural Language Generation Conference*. Edinburgh, UK, pages 265–273. arXiv:1608.00339. http://aclweb.org/anthology/W16-2302.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, PA, USA, pages 311–318. http://aclweb.org/anthology/P02-1040.

Laura Perez-Beltrachini and Claire Gardent. 2017. Analysing data-to-text generation benchmarks. In *Proceedings of the 10th International Natural Language Generation Conference*. Santiago de Compostela, Spain. http://arxiv.org/abs/1705.03802.

Verena Rieser and Oliver Lemon. 2009. Natural language generation as planning under uncertainty for spoken dialogue systems. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*. Athens, Greece, pages 683–691. http://aclweb.org/anthology/E09-1078.

Verena Rieser, Oliver Lemon, and Simon Keizer. 2014. Natural language generation as incremental planning under uncertainty: Adaptive information presentation for statistical dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22(5):979–993. https://doi.org/10.1109/TASL.2014.2315271.

Shikhar Sharma, Jing He, Kaheer Suleman, Hannes Schulz, and Philip Bachman. 2016. Natural language generation in dialogue using lexicalized and delexicalized data. *CoRR* abs/1606.03632. http://arxiv.org/abs/1606.03632.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*. Barcelona, Spain, pages 79–86. http://aclweb.org/anthology/P04-1011.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA, pages 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087.

Marilyn A Walker, Stephen J Whittaker, Amanda Stent, Preetam Maloor, Johanna Moore, Michael Johnston, and Gunaranjan Vasireddy. 2004. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science* 28(5):811–840. https://doi.org/10.1016/j.cogsci.2004.06.002.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Lina Maria Rojas-Barahona, Pei-hao Su, David Vandyke, and Steve J. Young. 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, pages 120–129. arXiv:1603.01232. http://aclweb.org/anthology/N16-1015.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 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, pages 1711–1721. http://aclweb.org/anthology/D15-1199.