Semantic Noise Matters for Neural Natural Language Generation

Ondřej Dušek*
Charles University
Faculty of Mathematics and Physics
Prague, Czech Republic
odusek@ufal.mff.cuni.cz

David M. Howcroft* & **Verena Rieser**
The Interaction Lab, MACS
Heriot-Watt University
Edinburgh, Scotland, UK
{d.howcroft,v.t.rieser}@hw.ac.uk

Abstract

Neural natural language generation (NNLG) systems are known for their pathological outputs, i.e. generating text which is unrelated to the input specification. In this paper, we show the impact of semantic noise on state-of-the-art NNLG models which implement different semantic control mechanisms. We find that cleaned data can improve semantic correctness by up to 97%, while maintaining fluency. We also find that the most common error is omitting information, rather than hallucination.

1 Introduction

Neural Natural Language Generation (NNLG) is promising for generating text from Meaning Representations (MRs) in an 'end-to-end' fashion, i.e. without needing alignments (Wen et al., 2015, 2016; Dušek and Jurčíček, 2016; Mei et al., 2016). However, NNLG requires large volumes of in-domain data, which is typically crowdsourced (e.g. Mairesse et al., 2010; Novikova et al., 2016; Wen et al., 2015, 2016; Howcroft et al., 2017), introducing noise. For example, up to 40% of the E2E Generation Challenge[1] data contains omitted or additional information (Dušek et al., 2019).

In this paper, we examine the impact of this type of semantic noise on two state-of-the-art NNLG models with different semantic control mechanisms: TGen (Dušek and Jurčíček, 2016) and SC-LSTM (Wen et al., 2015). In particular, we investigate the systems' ability to produce fact-accurate text, i.e. without omitting or hallucinating information, in the presence of semantic noise.[2] We find that:

- training on cleaned data reduces slot-error rate up to 97% on the original evaluation data;
- testing on cleaned data is challenging, even for models trained on cleaned data, likely due to increased MR diversity in the cleaned dataset; and
- TGen performs better than SC-LSTM, even when cleaner training data is available. We hypothesise that this is due to differences in how the two systems handle semantic input and the degree of delexicalization that they expect.

In addition, we release our code and a cleaned version of the E2E data with this paper.[3]

2 Mismatched Semantics in E2E Data

The E2E dataset contains input MRs and corresponding target human-authored textual references in the restaurant domain. MRs here are sets of attribute-value pairs (see Figure 1). Most MRs in the dataset have multiple references (8.1 on average). These were collected using crowdsourcing, leading to noise when crowd workers did not verbalise all attributes or added information not present in the MR. According to Dušek et al. (2019), the multiple references should help NLG systems abstract from the noise. However, most NLG systems in the E2E challenge in fact produced noisy outputs, suggesting that they were unable to learn to ignore noise in the training input.

Problems with the semantic accuracy in training data is not unique to the E2E dataset. Howcroft et al. (2017) collected a corpus of paraphrases differing with respect to information density for use in training NLG systems and found that subjects' paraphrases dropped about 5% of the slot-value pairs from the original texts and changed the val-

*Denotes equal contribution.

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

[2] Also see https://ehudreiter.com/2018/11/12/hallucination-in-neural-nlg/

[3] Data cleaning scripts, the resulting cleaned data and links to code are available at https://github.com/tuetschek/e2e-cleaning.

Proceedings of The 12th International Conference on Natural Language Generation, pages 421–426,
Tokyo, Japan, 28 Oct - 1 Nov, 2019. ©2019 Association for Computational Linguistics

Original MR: name[Cotto], eatType[coffee shop], food[English], priceRange[less than £20], customer_rating[low], area[riverside], near[The Portland Arms]

Human reference 1 (accurate): At the riverside near The Portland Arms, Cotto is a coffee shop that serves English food at less than £20 and has low customer rating.

HR 2: Located near The Portland Arms in riverside, the Cotto coffee shop serves English food with a price range of £20 and a low customer rating.
Corrected MR: name[Cotto], eatType[coffee shop], food[English], customer_rating[low], area[riverside], near[The Portland Arms]
(removed price range)

HR 3: Cotto is a coffee shop that serves English food in the city centre. They are located near the Portland Arms and are low rated.
Corrected MR: name[Cotto], eatType[coffee shop], food[English], customer_rating[low], area[city centre], near[The Portland Arms]
(removed price range, changed area)

HR 4: Cotto is a cheap coffee shop with one-star located near The Portland Arms.
Corrected MR: name[Cotto], eatType[coffee shop], priceRange[less than £20], customer rating[low], near[The Portland Arms]
(removed area)

Figure 1: MR and references from the E2E corpus. The first reference is accurate and verbalises all attributes, but the remaining ones contain inaccuracies. Corrected MRs were automatically produced by our slot matching script (see Section 3). Note that HR 2 is not fixed properly since the script's patterns are not perfect.

Dataset	Part	MRs	Refs	SER(%)
Original	TRAIN	4,862	42,061	17.69
	DEV	547	4,672	11.42
	TEST	630	4,693	11.49
Cleaned	TRAIN	8,362	33,525	(0.00)
	DEV	1,132	4,299	(0.00)
	TEST	1,358	4,693	(0.00)

Table 1: Data statistics comparison for the original E2E data and our cleaned version (number of distinct MRs, total number of textual references, SER as measured by our slot matching script, see Section 3).

ues for approximately 10% of the slot-value pairs. As a result of these changes and the insertion of new facts, only 61% of the corpus contained all and only the intended propositions. This is similar to what Eric et al. (2019) found in their work on the MultiWOZ 2.0 dataset: correcting the dialogue state annotations resulted in changes to about 40% of the dialogue turns in their dataset. These findings suggest that efforts to create more accurate training data—whether through stricter crowdsourcing protocols, conducting follow-up annotations (cf. Eric et al., 2019), or automated cleanup heuristics like we report here—are likely necessary in the NLG and dialogue systems communities.

3 Cleaning the Meaning Representations

To produce a cleaned version of the E2E data, we used the original human textual references, but paired them with correctly matching MRs.[4] To this end, we reimplemented the slot matching script of Reed et al. (2018), which tags MR slots and values using regular expressions. We tuned our expressions based on the first 500 instances from the E2E development set and ran the script on the full dataset, producing corrected MRs for all human references (see Figure 1). The differences against the original MRs allow us to compute the *semantic/slot error rate* (SER; Wen et al., 2015; Reed et al., 2018; Dušek et al., 2019):

$$\text{SER} = \frac{\#\text{added} + \#\text{missing} + \#\text{wrong value}}{\#\text{slots}}$$

To guarantee the integrity of the test set, we removed instances from the TRAIN (training) and DEV (development) sets that overlapped the TEST set. This resulted in 20% reduction for TRAIN and ca. 8% reduction for DEV in terms of references (see Table 1). On the other hand, the number of distinct MRs rose sharply after reannotation; the MRs also have more variance in the number of attributes. This means that the cleaned dataset is more complex overall, with fewer references per MR and more diverse MRs.

We manually evaluated 200 randomly chosen instances from the cleaned TRAIN set to check the accuracy of the slot matching script. We found that the slot matching script itself has a SER of 4.2%, with 39 instances (19.5%) not 100% correctly rated. This is much lower than the E2E dataset authors' own manual assessment of ca. 40% noisy instances (Dušek et al., 2019) and the script's rating of the whole dataset (mean SER: 16.37%),and comparable to the slot matching script of Juraska et al. (2018) evaluated on the same data.[5]

4 Evaluating the Impact on Neural NLG

We chose two recent neural end-to-end NLG systems, which represent two different approaches to semantic control and have been widely used and extended by the research community.

[4] Note that this can be done automatically, unlike fixing the references to match the original MRs.

[5] Juraska et al. (2018)'s script reaches 6.2% SER and 60 instances with errors, most of which is just omitting the *eatType[restaurant]* value. If we ignore this value, it gets 1.9% SER and 20 incorrect instances. We did not use this script as it was not available to us until very shortly before the camera-ready deadline. The script is now accessible under `https://github.com/jjuraska/slug2slug`. We plan to further improve our slot matching script based on errors found during the manual evaluation and comparison to Juraska et al. (2018).

4.1 TGen

TGen (Dušek and Jurčíček, 2016) is the baseline system used in the E2E challenge.[6] TGen is in essence a vanilla sequence-to-sequence (seq2seq) model with attention (Bahdanau et al., 2015) using LSTM cells where input MRs are encoded as sequences of triples in the form (dialogue act, slot, value).[7] TGen adds to the standard seq2seq setup a reranker that selects the output with the lowest SER from the decoder output beam (n-best list). SER is estimated based on a classifier trained to identify the MR corresponding to a given text. We use the default TGen parameters for the E2E data, experimenting with three variants:

- **TGen without reranker:** a vanilla seq2seq model with attention (TGen$-$);
- **TGen with default reranker:** the same augmented with an LSTM encoder and binary classifier for individual slot-value pairs;
- **TGen with oracle reranker:** directly uses the slot matching script to compute SER (TGen$+$).

We fixed the parameters of the main seq2seq generator to see the direct influence of each reranker, without the added effect of random initialization.

4.2 SC-LSTM

In contrast to seq2seq architecture used by TGen, the Semantically Controlled LSTM (SC-LSTM, Wen et al., 2015) uses a learned gating mechanism to selectively express parts of the MR during generation. We use the SC-LSTM model provided as part of the RNNLG repository[8] with minor changes to improve comparability to TGen. Most importantly, we incorporate the tokenization and normalization used by TGen into RNNLG. Since the word embeddings provided with RNNLG only cover about half of the tokens in the E2E dataset, we use randomly initialised word embeddings (dimension 50; same as TGen).

5 Evaluation and Results

To measure the effect of noisy data, we compare systems trained on the original data against systems trained using cleaned TRAIN and validation (=DEV) sets; we perform the comparisons both on the original and the cleaned TEST sets. Note that only scores on the same test set are directly comparable as the cleaned TEST set has more diverse MRs and fewer references per MR (i.e. numbers in Tables 2 and 3 cannot be compared across tables; cf. Section 3).

5.1 Automatic Metrics

We use freely available word-overlap-based evaluation metrics (WOM) scripts that come with the E2E data (Dušek et al., 2019),[9] supporting BLEU (Papineni et al., 2002), NIST (Doddington, 2002), ROUGE-L (Lin, 2004), METEOR (Lavie and Agarwal, 2007) and CIDEr (Vedantam et al., 2015). In addition, we use our slot matching script for SER (cf. Section 3). We also show detailed results for the percentages of added and missed slots and wrong slot values.[10]

The results in Table 2 (top half) for the original setup confirm that the ranking mechanism for TGen is effective for both WOMs and SER, whereas the SC-LSTM seems to have trouble scaling to the E2E dataset. We hypothesise that this is mainly due to the amount of delexicalisation required. However, the main improvement of SER comes from training on cleaned data with up to 97% error reduction with the ranker and 94% without.[11] In other words, just cleaning the training data has a much more dramatic effect than just using a semantic control mechanism, such as the reranker (0.97% vs. 4.27% SER). WOMs are slightly lower for TGen trained on the cleaned data, except for NIST, which gives more importance to matching less frequent n-grams. This suggests better preservation of content at the expense of slightly lower fluency.

The results for testing on cleaned data (Table 3, top half) confirm the positive impact of cleaned training data and also show that the cleaned test data is more challenging (cf. Section 3), as reflected in the lower WOMs. This raises the question whether the improved results from clean training data are due to seeing more challenging examples at training time. However, the improved results for training and testing on clean data (i.e. seeing equally challenging examples at training and test time), suggest the increase in performance can be attributed to data accuracy rather than diversity.

Looking at the detailed results for the number of

[6] https://github.com/UFAL-DSG/tgen

[7] The dialogue act is constant/ignored for the E2E dataset since it's not part of the MRs there.

[8] https://github.com/shawnwun/RNNLG

[9] https://github.com/tuetschek/e2e-metrics

[10] Absolute numbers of errors and number of completely correct instances are shown in Table 5 in the Supplementary.

[11] $\frac{0.12}{4.27} = 0.028$ and $\frac{0.97}{15.94} = 0.061$

Train	Test	System	BLEU	NIST	METEOR	ROUGE-L	CIDEr	Add	Miss	Wrong	SER
Original	Original	TGen−	63.37	7.7188	41.99	68.53	1.9355	00.06	15.77	00.11	15.94
		TGen	66.41	8.5565	45.07	69.17	2.2253	00.14	04.11	00.03	04.27
		TGen+	67.06	8.5871	45.83	69.73	2.2681	00.04	01.75	00.01	01.80
		SC-LSTM	39.11	5.6704	36.83	50.02	0.6045	02.79	18.90	09.79	31.51
Cleaned		TGen−	65.87	8.6400	44.20	67.51	2.1710	00.20	00.56	00.21	00.97
		TGen	66.24	8.6889	44.66	67.85	2.2181	00.10	00.02	00.00	00.12
		TGen+	65.97	8.6630	44.45	67.59	2.1855	00.02	00.00	00.00	00.03
		SC-LSTM	38.52	5.7125	37.45	48.50	0.4343	03.85	17.39	08.12	29.37
Cleaned missing		TGen−	66.28	8.5202	43.96	67.83	2.1375	00.14	02.26	00.22	02.61
		TGen	67.00	8.6889	44.97	68.19	2.2228	00.06	00.44	00.03	00.53
		TGen+	66.74	8.6649	44.84	67.95	2.2018	00.00	00.21	00.03	00.24
Cleaned added		TGen−	64.40	7.9692	42.81	68.87	2.0563	00.01	13.08	00.00	13.09
		TGen	66.23	8.5578	45.12	68.87	2.2548	00.04	03.04	00.00	03.09
		TGen+	65.96	8.5238	45.49	68.79	2.2456	00.00	01.44	00.00	01.45

Table 2: Results evaluated on the original test set (averaged over 5 runs with different random initialisation). See Section 5.1 for explanation of metrics. All numbers except NIST and ROUGE-L are percentages. Note that the numbers are *not* comparable to Table 3 as the test set is different.

Train	Test	System	BLEU	NIST	METEOR	ROUGE-L	CIDEr	Add	Miss	Wrong	SER
Original	Cleaned	TGen−	36.85	5.3782	35.14	55.01	1.6016	00.34	09.81	00.15	10.31
		TGen	39.23	6.0217	36.97	55.52	1.7623	00.40	03.59	00.07	04.05
		TGen+	40.25	6.1448	37.50	56.19	1.8181	00.21	01.99	00.05	02.24
		SC-LSTM	23.88	3.9310	32.11	39.90	0.5036	07.73	17.76	09.52	35.03
Cleaned		TGen−	40.19	6.0543	37.38	55.88	1.8104	00.17	01.31	00.25	01.72
		TGen	40.73	6.1711	37.76	56.09	1.8518	00.07	00.72	00.08	00.87
		TGen+	40.51	6.1226	37.61	55.98	1.8286	00.02	00.63	00.06	00.70
		SC-LSTM	23.66	3.9511	32.93	39.29	0.3855	07.89	15.60	08.44	31.94
Cleaned missing		TGen−	40.48	6.0269	37.26	56.19	1.7999	00.43	02.84	00.26	03.52
		TGen	41.57	6.2830	37.99	56.36	1.8849	00.37	01.40	00.09	01.86
		TGen+	41.56	6.2700	37.94	56.38	1.8827	00.21	01.04	00.07	01.31
Cleaned added		TGen−	35.99	5.0734	34.74	54.79	1.5259	00.02	11.58	00.02	11.62
		TGen	40.07	6.1243	37.45	55.81	1.8026	00.05	03.23	00.01	03.29
		TGen+	40.80	6.2197	37.86	56.13	1.8422	00.01	01.87	00.01	01.88

Table 3: Results evaluated on the cleaned test set (cf. Table 2 for column details; note that the numbers are *not* comparable to Table 2 as the test set is different).

Training data	Add	Miss	Wrong	Disfl
Original	0	22	0	14
Cleaned added	0	23	0	14
Cleaned missing	0	1	0	2
Cleaned	0	0	0	5

Table 4: Results of manual error analysis of TGen on a sample of 100 instances from the original test set: total absolute numbers of errors we found (added, missed, wrong values, slight disfluencies).

added, missing, and wrong-valued slots (Add, Miss, Wrong), we observe more deletions than insertions, i.e. the models more often fail to realise part of the MR, rather than hallucinating additional information. To investigate whether this effect stems from the training data, we partially cleaned the data of missing or added information only.[12] However, the results in bottom halves of Tables 2 and 3 do not support our hypothesis: we observe the main effect on SER from cleaning the missed slots, reducing both insertions and deletions. Again, one possible explanation is that cleaning the missing slots provided more complex training examples.

5.2 Manual Error Analysis

We carried out a detailed manual error analysis of selected systems to confirm the automatic metrics results, performing a blind annotation of semantic and fluency errors (not a human preference rating). We evaluated a sample of 100 outputs on the original test set produced by TGen with the default reranker trained using all four cleaning settings (original data, cleaned missing slots, cleaned added slots, fully cleaned). The results in Table 4 confirm the findings of the automatic metrics: systems trained on the fully cleaned set or the set with cleaned missing slots have near-perfect per-

[12] We only performed these experiments on TGen because of the low performance of SC-LSTM in general.

formance, with the fully-cleaned one showing a few more slight disfluencies than the other. The systems trained on the original data or with cleaned added slots clearly perform worse in terms of both semantic accuracy and fluency. All fluency problems we found were very slight and no added or wrong-valued slots were found, so missed slots are the main problem.

The manual error analysis also served to assess the accuracy of the SER measuring script on system outputs. Since NNLG tends to use more frequent phrasing, we expected better performance than on the dataset itself, and this proved true: we only found 2 errors in the 400 system outputs (i.e. 99.5% of instances and 99.93% of slots were matched correctly). This confirms that the automatic SER numbers reflect the semantic accuracy of individual systems very closely.

6 Discussion and Related Work

We present a detailed study of semantic errors in NNLG outputs and how these relate to noise in training data. We found that even imperfectly cleaned input data significantly improves semantic accuracy for seq2seq-based generators (up to 97% relative error reduction with the reranker), while only causing a slight decrease in fluency.

Contemporaneous with our work is the effort of Nie et al. (2019), who focus on automatic data cleaning using a NLU iteratively bootstrapped from the noisy data. Their analysis similarly finds that omissions are more common than hallucinations. Correcting for missing slots, i.e. forcing the generator to verbalise all slots during training, leads to the biggest performance improvement. This phenomenon is also observed by Dušek et al. (2018, 2019) for systems in the E2E NLG challenge, but stands in contrast to work on related tasks, which mostly reports on hallucinations (i.e. adding information not grounded in the input), as observed for image captioning (Rohrbach et al., 2018), sports report generation (Wiseman et al., 2017), machine translation (Koehn and Knowles, 2017; Lee et al., 2019), and question answering (Feng et al., 2018). These previous works suggest that the most likely case of hallucinations is an over-reliance on language priors, i.e. memorising 'which words go together'. Similar priors could equally exist in the E2E data for omitting a slot; this might be connected with the fact that the E2E test set MRs tend to be longer than training MRs (6.91 slots on average for test MRs vs. 5.52 for training MRs) and that a large part of them is 'saturated', i.e. contains all possible 8 attributes.

Furthermore, in accordance with our observations, related work also reports a relation between hallucinations and data diversity: Rohrbach et al. (2018) observe an increase for "novel compositions of objects at test time", i.e. non-overlapping test and training sets (cf. Section 3); whereas Lee et al. (2019) reports data augmentation as one of the most efficient counter measures. In future work, we plan to experimentally manipulate these factors to disentangle the relative contributions of data cleanliness and diversity.

Acknowledgments

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 authors would also like to thank Prof. Ehud Reiter, whose blog[13] inspired some of this research.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations (ICLR2015)*, San Diego, CA, USA. arXiv:1409.0473.

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*, pages 138–145, San Diego, CA, USA.

Ondřej Dušek and Filip Jurčíček. 2016. 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 (Volume 2: Short Papers)*, pages 45–51, Berlin, Germany. arXiv:1606.05491.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG Challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 322–328, Tilburg, The Netherlands. arXiv:1810.01170.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2019. Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Computer Speech & Language*, 59:123–156. arXiv:1901.07931.

[13] `https://ehudreiter.com/`

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. 2019. MultiWOZ 2.1: Multi-Domain Dialogue State Corrections and State Tracking Baselines. arXiv:1907.01669.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of Neural Models Make Interpretations Difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium.

David M. Howcroft, Dietrich Klakow, and Vera Demberg. 2017. The Extended SPaRKy Restaurant Corpus: Designing a Corpus with Variable Information Density. In *Proceedings of Interspeech 2017*, pages 3757–3761, Stockholm, Sweden.

Juraj Juraska, Panagiotis Karagiannis, Kevin K. Bowden, and Marilyn A. Walker. 2018. A Deep Ensemble Model with Slot Alignment for Sequence-to-Sequence Natural Language Generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 152–162, New Orleans, LA, USA.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada.

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*, pages 228–231, Prague, Czech Republic.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation. OpenReview.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 74–81, Barcelona, Spain.

F. Mairesse, M. Gašić, F. Jurčíček, S. Keizer, B. Thomson, K. Yu, and S. 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*, page 1552–1561, Uppsala, Sweden.

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*, pages 720–730, San Diego, CA, USA. arXiv:1509.00838.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics Volume 1: Long Papers*, pages 2673–2679, Florence, Italy.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing NLG Data: Pictures Elicit Better Data. In *The 9th International Natural Language Generation conference INLG*, Edinburgh, Scotland, UK. arXiv: 1608.00339.

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*, pages 311–318, Philadelphia, PA, USA.

Lena Reed, Shereen Oraby, and Marilyn Walker. 2018. Can Neural Generators for Dialogue Learn Sentence Planning and Discourse Structuring? In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 284–295, Tilburg, The Netherlands. arXiv:1809.03015.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium.

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)*, pages 4566–4575, Boston, MA, USA.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve 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*, pages 120–129, San Diego, CA, USA. arXiv: 1603.01232.

Tsung-Hsien Wen, Milica Gasic, 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*, pages 1711–1721, Lisbon, Portugal.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in Data-to-Document Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2243–2253, Copenhagen, Denmark. arXiv:1707.08052.