Designing a Symbolic Intermediate Representation for Neural Surface Realization

Henry Elder
ADAPT Centre
Dublin City University
henry.elder@adaptcentre.ie

Jennifer Foster
ADAPT Centre
Dublin City University
jennifer.foster@dcu.ie

James Barry
ADAPT Centre
Dublin City University
james.barry@adaptcentre.ie

Alexander O'Connor
Autodesk, inc.
alex.oconnor@autodesk.com

Abstract

Generated output from neural NLG systems often contain errors such as hallucination, repetition or contradiction. This work focuses on designing a symbolic intermediate representation to be used in multi-stage neural generation with the intention of reducing the frequency of failed outputs. We show that surface realization from this intermediate representation is of high quality and when the full system is applied to the E2E dataset it outperforms the winner of the E2E challenge. Furthermore, by breaking out the surface realization step from typically end-to-end neural systems, we also provide a framework for non-neural content selection and planning systems to potentially take advantage of semi-supervised pretraining of neural surface realization models.

1 Introduction

For Natural Language Generation (NLG) systems to be useful in practice, they must generate utterances that are adequate, that is, the utterances need to include all relevant information. Furthermore the information should be expressed correctly and fluently, as if written by a human. The rule and template based systems which dominate commercial NLG systems are limited in their generation capabilities and require much human effort to create but are reliably adequate and known for widespread usage in areas such as financial journalism and business intelligence. By contrast, neural NLG systems need only a well collected dataset to train their models and generate fluent sounding utterances but have notable problems, such as hallucination and a general lack of adequacy (Wiseman et al., 2017). There was a marked absence of neural NLG in any of the finalist systems in either the 2017 or 2018 Alexa Prize (Fang et al., 2017; Chen et al., 2018).

Following prior work in the area of multi-stage neural NLG (Dušek and Jurcicek, 2016; Daniele et al., 2017; Puduppully et al., 2018; Hajdik et al., 2019; Moryossef et al., 2019), and inspired by more traditional pipeline data-to-text generation (Reiter and Dale, 2000; Gatt and Krahmer, 2018), we present a system which splits apart the typically end-to-end data-driven neural model into separate utterance planning and surface realization models using a symbolic intermediate representation. We focus in particular on surface realization and introduce a new symbolic intermediate representation which is based on an underspecified universal dependency tree (Mille et al., 2018b). In designing our intermediate representation, we are driven by the following constraints:

1. The intermediate representation must be suitable for processing with a neural system.
2. It must not make the surface realization task too difficult because we are interested in understanding the limitations of neural generation even under favorable conditions.
3. It must be possible to parse a sentence into this representation so that a surface realization training set can be easily augmented with additional in-domain data.

Focusing on English and using the E2E dataset, we parse the reference sentences into our intermediate representation. We then train a surface realization model to generate from this representation, comparing the resulting strings with the reference using both automatic and manual evaluation. We find that the quality of the generated text is high, achieving a BLEU score of 82.47. This increases to 83.38 when we augment the training data with sentences from the TripAdvisor corpus. A manual error analysis shows that in only a very small proportion (∼5%) of the output sentences, the meaning of the reference is not fully recovered. This

Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen), pages 65–73
Minneapolis, Minnesota, USA, June 6, 2019.

high level of adequacy is expected since the intermediate representations are generated directly from the reference sentences. An analysis of a sample of the adequate sentences shows that readability is on a par with the reference sentences.

Having established that surface realization from our intermediate representation achieves sufficiently high performance, we then test its efficacy as part of a pipeline system. On the E2E task, our system scores higher on automated results than the winner of the E2E challenge (Juraska et al., 2018). The use of additional training data in the surface realization stage results in further gains. These encouraging results suggest that pipelines can work well in the context of neural NLG.

2 Methods

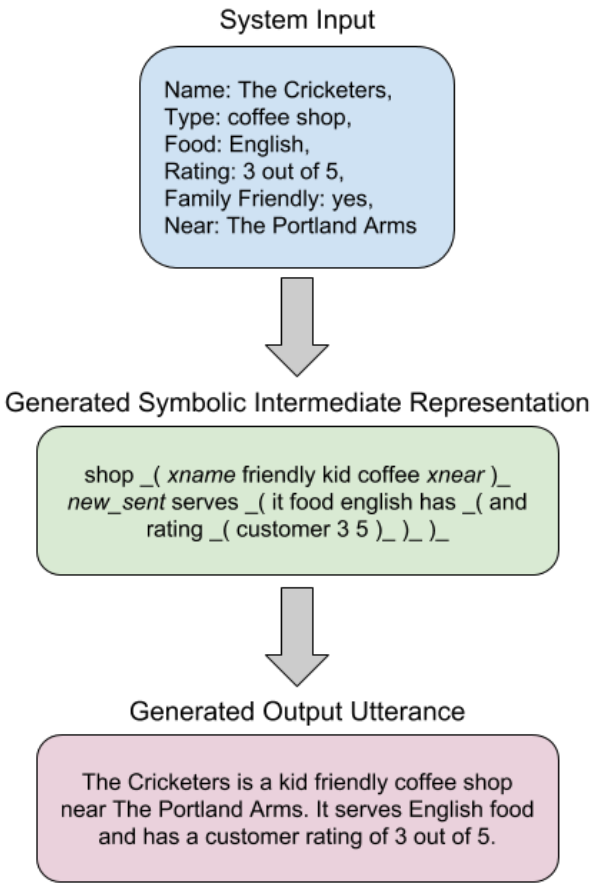

Figure 1: Example of two stage generation using the pipeline system. Both are real examples generated by their respective models.

Our system consists of two distinct models. The first is an utterance planning model which takes as input some structured data and generates an intermediate representation of an utterance containing one or more sentences. The intermediate representation of each sentence in the utterance is then passed to a second surface realization model which generates the final natural language text. See Figure 1 for an example from the E2E dataset. Both models are neural based. We use a symbolic intermediate representation to pass information between the two models.

2.1 Symbolic Intermediate Representation

The symbolic intermediate representation used is the *deep*[1] Underspecified Universal Dependency (UUD) structure (Mille et al., 2018b). The UUD structure is a tree "containing only content words linked by predicate-argument edges in the PropBank/NomBank (Palmer et al., 2005; Meyers et al., 2004) fashion" (Mille et al., 2018b). Each UUD structure represents a single sentence. The UUD structure was designed to "approximate the kind of abstract meaning representations used in native NLG tasks" (Mille et al., 2018b). That is, the kind of output that a rule based system could be reasonably expected to generate as part of a pipeline NLG process. However, to the best of our knowledge, no such system has yet been developed or adapted to generate the deep UUD structure as output. Hence it was required to make a number of changes to the deep UUD structure during preprocessing to better suit a neural system designed to use the structure as a symbolic intermediate representation; namely we linearize the UUD tree, remove accompanying token features and use the surface form of each token, see Figure 2.

Linearization In order to use tree structures in a sequence-to-sequence model a linearization order for nodes in the tree must be determined. Following Konstas et al. (2017) tree nodes are ordered using depth first search. Scope markers are added before each child node. When a node has only one child node we omit scope markers. Though this can lead to moderate ambiguity it greatly reduces the length of the sequence (Konstas et al., 2017).

When two nodes appear at the same level in the tree their linearization order is typically chosen at random, or using some rule based heuristic or even a secondary model (Ferreira et al., 2018). In this system linearization of equivalent level tokens is

[1] *deep* here is not referring to deep learning but rather as a contrast with another UUD variant known as the *shallow* UUD. Shallow and deep surface realization tracks were used in both Surface Realization Shared Tasks (Mille et al., 2018a; Belz et al., 2011)

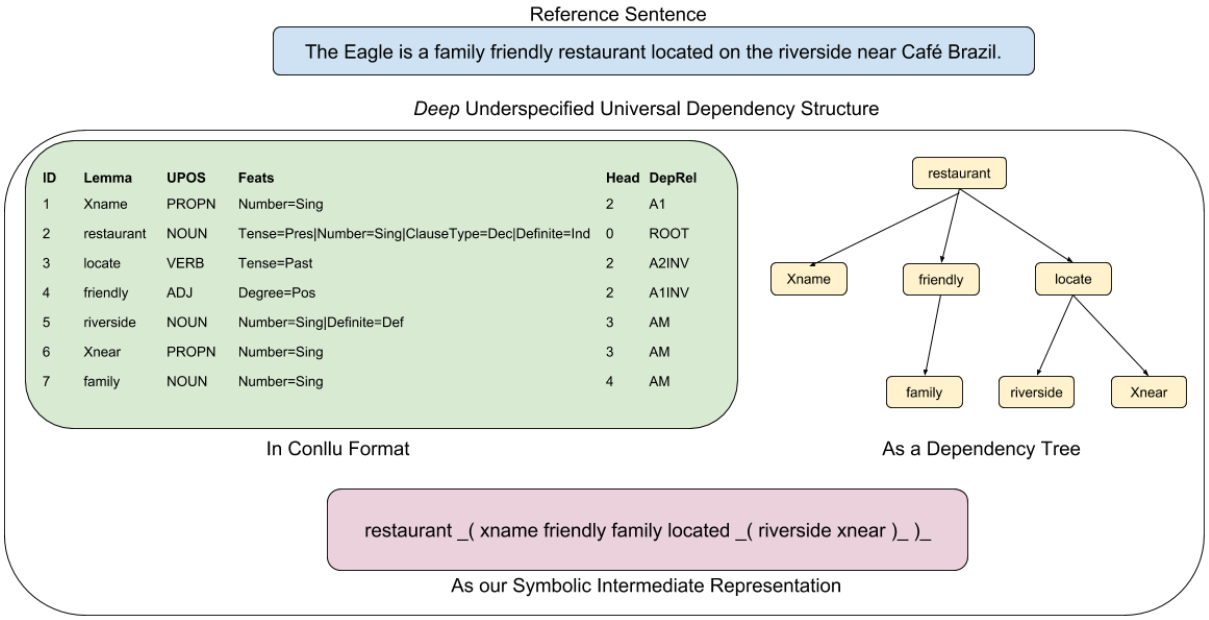

Figure 2: Different representations of the *Deep* Underspecified Universal Dependency structure

determined by the original order in which they appeared in the sentence. We chose to use a consistent, as opposed to random, ordering of equivalent level nodes for the symbolic intermediate representation as it has been shown in a number of papers (Konstas et al., 2017; Juraska et al., 2018) that neural models perform worse at given tasks when trained on symbolic intermediate representations sorted in random orders, even when that randomness is used to augment and increase the size of the data. We chose to use original sentence order of tokens as the basis for ordering sibling nodes. Though this is clearly a simplification, and gives the model additional information, it is an intuitive choice.

Features As well as the head id, tokens in the deep UUD structure are each associated with a number of additional features: dependency relations (DepRel), universal part-of-speech tag (UPOS) and lexical features (feats), see Figure 2. Other neural based work on surface realization from the deep UUD structure included this information using factor methods (Elder and Hokamp, 2018). However our symbolic intermediate representation does not include these additional features. By not including the additional features with each token we simplify the task of generating the symbolic intermediate representation using a neural model. Token features could be generated using multitask learning as in Dalvi et al. (2017) but we leave this for future work.

Lemmas vs. Forms In the deep UUD structure the token provided is a lemma, the root of the original form of a token. Part-of-speech and lexical features are provided to enable a surface realization system to determine the form. As we do not include these features in our symbolic intermediary representation we use the original form of token instead. This is another simplification of the surface realization task. While we found that *lemma + part of speech tag + lexical features* typically provide enough information to reconstruct the original form, it is not a 100% accurate mapping.

3 Experiments

Datasets Experiments were performed with the E2E dataset (Novikova et al., 2017). Figure 1 contains an example of the E2E input. The E2E dataset contains a training set of 42,061 pairs of meaning representations and utterances. Training data for the surface realization model was augmented, for some experiments, with the TripAdvisor corpus (Wang et al., 2010), which was filtered for sentences with a 100% vocabulary overlap with the E2E corpus and a sentence length between 5 and 30 tokens, resulting in an additional 209,823 sentences, with an average sentence length of 10 tokens. By comparison the E2E corpus had sentence lengths ranging between 1 and

59 tokens with an average sentence length of 13 tokens.

Both corpora were sentence tokenized and parsed by the Stanford NLP universal dependency parser (Qi et al., 2018). The parsed sentences in CoNLL-U format were then further processed by a special deep UUD parser (Mille et al., 2018b). Utterances from the E2E corpus were delexicalised to anonymize restaurant names in both the *name* and *near* slots of the meaning representation. All tokens were lower cased before training.

Models For the neural NLG pipeline system we train two separate encoder-decoder models using the neural machine translation framework OpenNMT (Klein et al., 2017). We trained two separate encoder-decoder models for surface realization and content selection. However both used the same hyperparameters. A single layer LSTM (Hochreiter and Schmidhuber, 1997) with RNN size 450 and word vector size 300 was used. The models were trained using ADAM (Kingma and Ba, 2015) with a learning rate of 0.001. The only difference between the two models was that the surface realization model was trained with a copy attention mechanism (Vinyals et al., 2015).

For the full E2E task a single planning model was trained on the E2E corpus. However two different surface realization models were compared; one trained solely on sentences from the E2E corpus and another trained on a combined corpus of E2E and TripAdvisor sentences. For baselines on the full E2E task we compare with two encoder-decoder models which both use semantic rerankers on their generated utterances; TGen (Dušek and Jurcicek, 2016) the baseline system for the E2E challenge and Slug2Slug (Juraska et al., 2018) the winning system of the E2E challenge.

Automated Evaluation The E2E task is evaluated using an array of automated metrics[2]; BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Lavie and Agarwal, 2007), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015). The two surface realization models were evaluated on how well they were able to realize sentences from the E2E validation set using silver parsed intermediate representations. We report BLEU-4 scores[3] for the silver parse generated texts from the surface realization models. In both the E2E (Dušek et al., 2019) and WebNLG challenge (Shimorina, 2018) it was found that automated results did not correlate with the human evaluation *on the sentence level*. However in the Surface Realization shared task correlation between BLEU score and human evaluation was noted to be highly significant (Mille et al., 2018a).

Manual Analysis The importance of using human evaluation to get a more accurate understanding of the quality of text generated by an NLG system cannot be overstated. We perform human evaluation on the outputs of the surface realization model with a silver parse of the original utterances as input. We evaluate the outputs first in terms of meaning similarity and then readability and fluency.

To evaluate the surface realization model we compare generated utterances with the human references. For the meaning similarity human evaluation we remove sentences with no differences, only differences involving the presence or absence of hyphens or only capitalization differences. We evaluate meaning similarity between two utterances as whether they contain the same meaning. We treat this a binary Yes / No decision as the generated utterances are using a silver parse and ought to be able to reconstruct a sentence that, while possibly differently structured, does express the same meaning.

We manually analyze failure cases where semantic similarity is not achieved to discover where the issues arise. There may be failures in the method of obtaining the intermediate representation, in the surface realization model or some other issue with the intermediate representation.

We then pass on only those generated utterances deemed to have the same meaning with the reference utterance into the next stage of readability evaluation. To evaluate readability we perform pairwise comparisons between generated utterances and reference utterances. We randomize the order during evaluation so it is not clear what the origin of a particular utterance is. We define readability, sometimes called fluency, as how well a given utterances reads, "is it fluent English

[2]E2E NLG Challenge provides an official scoring script https://github.com/tuetschek/e2e-metrics

[3]We input tokenized, lowercased and relexicalised sentences to the Moses multi-bleu perl script: https://github.com/OpenNMT/OpenNMT-py/blob/master/tools/multi-bleu.perl

or does it have grammatical errors, awkward constructions, etc." (Mille et al., 2018a). By investigating readability of utterances with meaning similarity, we hope to see how the surface realization model performs compared with a human written utterance. The surface realization model is required to at least match human level performance in order to be usable, if it does not then we need to investigate where it fails and why. We used Prodigy (Montani and Honnibal, 2018) as our data annotation tool.

4 Results

4.1 Surface Realization Analysis

	BLEU
E2E	0.8247
+ TripAdvisor	**0.8338**

Table 1: Automated evaluation of surface realization models on validation set sentences

Automated evaluation To initially establish if training on additional data from a different corpus was beneficial we performed automated evaluation. Each surface realization model is provided a parse of the target sentence. The BLEU score is slightly higher, see Table 1, when the model is trained with the additional corpus data.

	E2E	+ TripAdvisor
Exact matches	3807	3935
Punctuation and/or determiner differences	1242	1268

Table 2: Surface Realization of 8024 sentences in the E2E validation set

	E2E	+ TripAdvisor
Remaining sentences	2975	2821
Sentences analysed	325	325
Failed meaning similarities	76	45
Same readability as reference	198	208
Worse readability than reference	30	43
Better readability than reference	21	29

Table 3: Manual analysis of a subset of remaining sentences from the 8024 sentences in the E2E validation set

Manual analysis Starting with generated sentences from the E2E validation set, we first filter out exact or very close matches to the reference sentences, see Table 2. Then taking a subset of remaining generated sentences, we establish that they contain the same meaning as the reference sentence. Finally we compare the readability / naturalness of the generated text with the human reference sentences, see Table 3.

While the surface realization model trained on both E2E and Trip Advisor corpora generally outperforms the model trained on only E2E data, it has more sentences rated as *Worse readability than reference*. More detailed manual analysis is required to tell whether this is a statistical anomaly or a true insight into how the additional data is affecting model performance.

Analysis of failed meaning similarities Looking at examples where a generated sentence failed to correctly capture the meaning of the reference sentence we find the causes for this fall into a number of categories:

- Poor sentence tokenization
- Problems with the reference sentence
- Unusually phrased reference sentence
- Unknown words
- Generation model failures (repetition or missing words)

The model trained on the additional TripAdvisor corpus has a larger vocabulary and has seen a wider range of sentences, and thus fails less often. Most failures appear to be due to reference sentences containing unknown tokens or being phrased in a new or unusual way the model has not seen before. A smaller number of cases are attributable to issues directly with the generation model, namely repetition or absence of tokens from the intermediate representation. Figure 3 contains three examples of failed generation.

	BLEU	NIST	METEOR	ROUGE_L	CIDEr
Validation					
TGen	0.6925	8.4781	0.4703	0.7257	2.3987
Slug2Slug	0.6576	8.0761	0.4675	0.7029	-
Pipeline	0.7271	8.5680	0.4874	0.7546	2.5481
+ TripAdvisor	**0.7298**	**8.5891**	**0.4875**	**0.7557**	**2.5507**
Test					
TGen	0.6593	8.6094	0.4483	0.6850	2.2338
Slug2Slug	0.6619	8.6130	0.4454	0.6772	2.2615
Pipeline	0.6705	8.6737	**0.4573**	0.7114	2.2940
+ TripAdvisor	**0.6738**	**8.7277**	0.4572	**0.7152**	**2.2995**

Table 4: Automated results on end-to-end task

Ref: Do not go to The Punter near riverside.

IR: go _(not xname riverside)_

Gen: Not go to The Punter in riverside.

(a) Model generation failure

Ref: With only an average customer rating, and it being a no for families, it doesn't have much going for it.

IR: have _(rating _(only average customer no _(and it families)_)_it n't much _(going it)_)_

Gen: With a only average customer rating and its no families, it won't have much that going to it.

(b) Unusual phrasing in reference sentence

Ref: Have you heard of The Sorrento and The Wrestlers, they are the average friendly families.

IR: heard _(you xnear _(xname and)_families _(they average friendly)_)_

Gen: You can be heard near The Sorrento and The Wrestlers, they are average friendly families.

(c) Nonsensical reference sentence

Figure 3: Examples of reference sentences *(Ref)*, intermediate representations *(IR)* and generated texts *(Gen)* from three different scenarios.

4.2 End-to-End Analysis

We report results on the full E2E task in Table 4. Both our systems outperform the E2E challenge winning system Slug2Slug (Juraska et al., 2018), with the system using the surface realization model trained with additional data performing slightly better. Both surface realization models received the same set of intermediate representations from the single utterance planning model.

Further human evaluation may be required to establish the meaningfulness of these higher automated results.

5 Related Work

The work most similar to ours is (Dušek and Jurcicek, 2016). It is also in the domain of task oriented dialogue and they apply two-stage generation; first generating deep syntax dependency trees using a neural model and then generating the final utterance using a non-neural surface realizer. They found that while generation quality is initially higher from the two-stage model, when using a semantic reranker it is outperformed by an end-to-end seq2seq system.

Concurrent to this work is Moryossef et al. (2019). In this work they split apart the task of planning and surface realization. Conversely to Dušek and Jurcicek (2016) they employ a rule based utterance planner and a neural based surface realizer. They applied this system to the WebNLG corpus (Gardent et al., 2017) and found that, compared with a strong neural system, it performed roughly equally at surface realization but exceeded the neural system at adequately including information in the generated utterance.

Other work has looked for innovative ways to separate planning and surface realization from the end-to-end neural systems, most notably Wiseman et al. (2018) which learns template generation also on the E2E task, but does not yet match baseline performance, and He et al. (2018) which has a dialogue manager control decision making and passes this information onto a secondary language generator. Other work has attempted either multi-stage semi-unconstrained language generation, such as in the domain of story telling (Fan et al., 2019), or filling-in-the-blanks style sentence reconstruction (Fedus et al., 2018).

6 Discussion

Our system's automated results on the E2E task exceed that of the winning system. This shows that splitting apart utterance planning and surface realization in a fully neural system may have potential benefit. Our intuition is that by loosely separating the semantic and syntactic tasks of sen-

tence planning and surface realization, our models are more easily able to learn alignments between source and target sequences in each distinct task than in a single model. More clear alignments may help as the E2E corpus is a relatively small dataset, at least compared with dataset sizes used for neural machine translation (Bojar et al., 2018) for which end-to-end neural models are the dominant paradigm. Further human analysis of the generated utterances' fluency and adequacy[4] could help determine what is driving the improved performance on automated metrics.

The design of our symbolic intermediate representation is such that additional training data can be easily collected for the surface realization model. Indeed we see marginally better results on the E2E task with a surface realization model trained on both the E2E and TripAdvisor corpuses. This approach could be further scaled beyond the relatively small number of additional sentences we automatically parsed from the TripAdvisor corpus. In the E2E challenge it was noted that a semantic reranker was requisite for high performing neural systems (Dušek et al., 2019). Adding a semantic reranker to our system could likely help improve performance of the utterance planning step.

While we made simplifications to the intermediate representation, namely including forms over lemmas and using the original sentence order to sort adjacent nodes, their generation was still required to be performed by a higher level model. It's possible that different higher level systems, for example a rule based utterance planning system, might prefer a more abstract intermediate representation. Indeed this trade off between what information ought to go into the intermediate representation is a highly practical one. A surface realization model trained using our automated representation could be made to work with a rule based system providing input.

7 Conclusion

We have designed a symbolic intermediate representation for use in a pipeline neural NLG system. We found the surface realization from this representation to be of high quality, and that results improved further when trained on additional data. When testing the full pipeline system automated results exceeded that of prior top performing neural systems, demonstrating the potential of breaking apart typically end-to-end neural systems into separate task-focused models.

[4]The generated utterance's coverage of the input meaning representation

Acknowledgements

This research is supported by Science Foundation Ireland in the ADAPT Centre for Digital Content Technology. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Anja Belz, Michael White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. 2011. The First Surface Realisation Shared Task: Overview and Evaluation Results. In *Proceedings of the European Workshop on Natural Language Generation*, December, pages 217–226.

Ondej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Chun-Yen Chen, Dian Yu, Weiming Wen, Yi Mang Yang, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, Shreenath Iyer, Giritheja Sreenivasulu, Runxiang Cheng, Ashwin Bhandare, and Zhou Yu. 2018. Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel. 2017. Understanding and Improving Morphological Learning in the Neural Machine Translation Decoder. *IJCNLP*, pages 142–151.

Andrea F Daniele, Matthew R Walter, Mohit Bansal, and Matthew R Walter. 2017. Navigational Instruction Generation as Inverse Reinforcement Learning with Neural Machine Translation. In *Proceedings of HRI*.

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. ®.

Ondrej Dušek and Filip Jurcicek. 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. Association for Computational Linguistics.

Ondej Dušek, Jekaterina Novikova, and Verena Rieser. 2019. Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Under Review*.

Henry Elder and Chris Hokamp. 2018. Generating High-Quality Surface Realizations Using Data Augmentation and Factored Sequence Models. In *Proceedings of the 1st Workshop on Multilingual Surface Realisation (MSR), 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for Structuring Story Generation. In *Forthcoming NAACL*.

Hao Fang, Hao Cheng, Elizabeth Clark, Ariel Holtzman, Maarten Sap, Mari Ostendorf, Yejin Choi, and Noah A Smith. 2017. Sounding Board University of Washingtons Alexa Prize Submission. In *1st Proceedings of Alexa Prize*, page 12.

William Fedus, Ian Goodfellow, and Andrew M Dai. 2018. MaskGAN: Better Text Generation via Filling in the_____. In *International Conference on Learning Representations*.

Thiago Castro Ferreira, Sander Wubben, Emiel Krahmer, Thiago Castro Ferreira, Sander Wubben, and Emiel Krahmer. 2018. Surface Realization Shared Task 2018 (SR18): The Tilburg University Approach. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, volume 2018, pages 35–38. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG Challenge: Generating Text from RDF Data. In *Proceedings of The 10th International Natural Language Generation conference*, pages 124–133.

Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61(c):65–170.

Valerie Hajdik, Jan Buys, Michael W Goodman, and Emily M Bender. 2019. Neural Text Generation from Rich Semantic Representations. In *Forthcoming NAACL*.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling Strategy and Generation in Negotiation Dialogues. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Juraj Juraska, Panagiotis Karagiannis, Kevin Bowden, and Marilyn 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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-Sequence Models for Parsing and Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Stroudsburg, PA, USA. Association for Computational Linguistics.

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*, StatMT '07, pages 228–231, Stroudsburg, PA, USA.

C Y Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the workshop on text summarization branches out (WAS 2004)*, 1, pages 25–26.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: An interim report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation, Boston, MA, May 2004*, pages 24–31.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018a. The First Multilingual Surface Realisation Shared Task (SR'18): Overview and Evaluation Results. In *Proceedings of the 1st Workshop on Multilingual Surface Realisation (MSR), 56th Annual Meeting of the Association for Computational Linguistics*, pages 1–10, Melbourne, Australia.

Simon Mille, Anja Belz, Bernd Bohnet, and Leo Wanner. 2018b. Underspecified Universal Dependency Structures as Inputs for Multilingual Surface Realisation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 199–209. Association for Computational Linguistics.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*, to appear.

Amit Moryossef, Yoav Goldberg, Ido Dagan, and Ramat Gan. 2019. Separating Planning from Realization in Neural Data to Text Generation. In *Forthcoming NAACL*.

Jekaterina Novikova, Ondej Dušek, and Verena Rieser. 2017. The E2E Dataset: New Challenges For End-to-End Generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, August, pages 201–206, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2018. Data-to-Text Generation with Content Selection and Planning. *Aaai*.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D Manning. 2018. Universal Dependency Parsing from Scratch. In *Proceedings of the (CoNLL) 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Anastasia Shimorina. 2018. Human vs Automatic Metrics: on the Importance of Correlation Design.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:4566–4575.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer Networks. In *Advances in Neural Information Processing Systems 28*, pages 2692–2700.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*, pages 783–792, Washington, DC, USA. ACM Press.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in Data-to-Document Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. Learning Neural Templates for Text Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187.