TNT-NLG, System 1: Using a Statistical NLG to Massively Augment Crowd-Sourced Data for Neural Generation

Shereen Oraby[1], **Lena Reed**[1], **Shubhangi Tandon**[1],
Sharath T.S.[1], **Stephanie Lukin**[2], **and Marilyn Walker**[1]
[1] University of California, Santa Cruz
[2] US Army Research Laboratory
{soraby,lireed,shtandon,sturuvek,mawalker}@ucsc.edu
stephanie.m.lukin.civ@mail.mil

Abstract

Ever since the successful application of sequence to sequence learning for neural machine translation systems (Sutskever et al., 2014), interest has surged in its applicability towards language generation in other problem domains. In the area of natural language generation (NLG), there has been a great deal of interest in end-to-end (E2E) neural models that learn and generate natural language sentence realizations in one step. In this paper, we present TNT-NLG System 1, our first system submission to the E2E NLG Challenge, where we generate natural language (NL) realizations from meaning representations (MRs) in the restaurant domain by massively expanding the training dataset. We develop two models for this system, based on Dusek et al.'s (2016a) open source baseline model and context-aware neural language generator. Starting with the MR and NL pairs from the E2E generation challenge dataset, we explode the size of the training set using PERSONAGE (Mairesse and Walker, 2010), a statistical generator able to produce varied realizations from MRs, and use our expanded data as contextual input into our models. We present evaluation results using automated and human evaluation metrics, and describe directions for future work.

1 Introduction

The growing popularity of conversational agents has sparked increased interest in natural language response generation (NLG). To this end, there has been a movement towards end-to-end (E2E) neural models that learn and generate sentence realizations in one step (Wen et al., 2015; Mei et al., 2016; Dušek and Jurcicek, 2016b; Lampouras and Vlachos, 2016).

The E2E NLG Challenge (Novikova et al., 2017) focuses on the specific task of using neural models to generate a natural language (NL) utterance given a meaning representation (MR). The requirements are that the generated utterance is both similar to an utterance that is written by a human (as a reference text), and also highly rated by humans for quality and naturalness.

In this paper, we describe our first system submission to the E2E NLG challenge, where we develop a model based on the open source context-aware neural language generator (Dušek and Jurcicek, 2016a). Instead of using the previous sentence to maintain the context of the conversation, which is the traditional way a context encoder is used, we experiment with providing a prior to the model in terms of what it should generate. Specifically, we explode the size of the training data available to the neural generator by generating synthetic training examples using PERSONAGE (Mairesse and Walker, 2008), a statistical language generator that is able to generate varied realizations from a given MR.

We present two models for our system, a primary model which uses context from PERSONAGE, and a secondary model which uses the same context but is trained on a more expanded dataset that we generate by repeating instances form training[1]. We present descriptions of the expanded training data and models, and show that our models outperform the baseline for the METEOR metric, and score competitively on the qualitative metrics of quality and naturalness.

2 Related Work

The restaurant domain has long been a testbed for conversational agents with considerable previous work on NLG (Howcroft et al., 2013; Polifroni et al., 1992; Whittaker et al., 2002; Stent et al., 2004; Devillers et al., 2004; Gašic et al., 2008;

[1] More detail about this method is presented in a companion paper for our second E2E submission, "TNT-NLG, System 2".

Mairesse et al., 2010). Until recently, NLG methods used a two step process of sentence planning followed by surface realization (Reiter and Dale, 2000), with stylistic variations in utterances controlled by the sentence planner (Stent et al., 2004).

The task of response generation for conversational agents has gradually evolved from a simple retrieval based method using similarity metrics, to a sequence to sequence generation problem (Serban et al., 2016; Vinyals and Le, 2015), and as well as an ensemble of both (Song et al., 2016). Retrieval based methods, while effective, are limited in their capacity because they only produce responses that have been seen before. Though an advantage of these models is that they do not produce output that is syntactically incorrect. A generative model on the other hand learns a language model, which is a distribution on words in the vocabulary conditioned on the previous ones, and is capable of producing sentences not seen in the training set.

This is one of the main reasons why the state of the art for machine translation cannot be directly transfered to the task of coherent response generation. A particular utterance can map to multiple coherent responses and the conventional loss functions, such as cross entropy, are not well suited for this task. Such loss functions make the model expect a particular response and penalize any others even though they are coherent in terms of semantics and context.

There has been focus on incorporating the context of the conversation in order to generate responses (Dušek and Jurcicek, 2016a; Serban et al., 2017). Sordoni et al. (2015) propose a simple approach for dialogue generation, where they incorporate the previous set of responses in a dialogue as a bag of words model and use a feed forward neural network to inject a fixed sized context vector into the LSTM cell of the encoder. Ghosh et al. (2016) proposed a modified LSTM cell with an additional gate that incorporates the previous context as input during encoding. The weights of the gate are learned exactly in the same way as the weights for the input, and forget and output gates are learned.

In this work, we propose a context model where a recurrent neural network is used to encode output from PERSONAGE as a prior, and decode with attention (Bahdanau et al., 2015). It also has the reranking classifier and ngram re-ranker based on BLEU score similarity between the output of the reranking classifier with input MRs, as proposed by Dusek et al. (Dušek and Jurcicek, 2016a,b).

3 Data

In the past, E2E NLG datasets were small and delexicalised, e.g BAGEL or RoboCup (Mairesse et al., 2010; Kitano et al., 1998). The data from the E2E NLG challenge consists of 50k crowd-sourced phrases in the restaurant domain (Novikova et al., 2016). This data is characterized by an open vocabulary, complex syntactic structures and diverse discourse phenomena.

The input is in the form of meaning representation (MR), which include dialogue attribute and value pairs. There are 8 different attributes, each with at least 2 values. The MRs have on average 8.1 corresponding crowd-sourced natural language realizations (references) (Novikova et al., 2016). All the attributes and example values are listed below:

- NAME=COCUM
- AREA=CITY CENTRE
- FOOD=ENGLISH
- EAT_TYPE=COFFEE SHOP
- FAMILY_FRIENDLY=YES
- NEAR=RAJA INDIAN CUISINE
- PRICE_RANGE=£20-25
- RATING=HIGH

The dataset consists of a train, dev and test set. The train set has 42k references and 4862 MRs, the dev set has 4672 references and 547 MRs, and the test set has 630 MRs.

3.1 Data Augmentation with Personage

We automatically generated sentences to augment the data using PERSONAGE, an NLG engine that performs content planning (Mairesse and Walker, 2008). PERSONAGE uses Deep Syntactic Structures (DSYNTS) which are syntactic-semantic representations of entries in PERSONAGE (Mel'cuk, 1988; Lavoie and Rambow, 1997). PERSONAGE controls the content size and structure of arguments, using content planning operations that are parameterized to make planning decisions including aggregation and lexical choice. For each attribute in the E2E data, we created a DSYNTS. PERSONAGE that can then aggregate the DSYNTS together to create restaurant descriptions using multiple attributes. The DSYNTS were created by hand, however we believe the effort justified because the amount of data that can be generated by

Attribute Type	DSYNT Name	Value	Sentence
EATTYPE	eattype	coffee shop	name_var is a *coffee shop*.
FOOD	cuisine	Italian	name_var is an *Italian* [restaurant/place].
PRICERANGE	price	cheap	name_var is *cheap*.
	pricecost	less than £20	name_var costs *less than £20*.
	pricerange	£20-25	name_var has a price range of *£20-25*.
CUSTOMER RATING	rating	3 out of 5	name_var has a *3 out of 5* rating.
AREA	area	city centre	name_var is in *city centre*.
FAMILYFRIENDLY	friendly	yes	name_var is *kid friendly*.
	notfriendly	no	name_var isn't *family friendly*.
NEAR	The Portland Arms	near	name_var is near *The Portland Arms*.

Table 1: PERSONAGE outputs for attribute types

a few small DSYNTS is very large. The DSYNTS created to represent the attributes, along with example delexicalized sentences, are in Table 1.

We take the MR for each sentence in the training data create a PERSONAGE textplan automatically. A textplan contains the DSYNTS that we will be used in the generated phrase and how these DSYNTS will be aggregated. The PERSONAGE outputs for a given MR were always unique, introducing more variability into our data. However, while PERSONAGE has a multitude of parameters that can be inserted, to mimic the E2E data, we did not include most of these parameters in our data. The attributes are combined into a realization using different aggregation operations. These aggregation operations are:

- PERIOD
- "WITH" CUE WORD
- CONJUNCTION (*X is Y and it is Z.*)
- MERGE (*X is Y and Z.*)
- "ALSO" CUE WORD
- ELLIPSIS

The system will also insert words such as "well" automatically to make the sentence sound more natural and to create more varied the data. In order to duplicate the exact number of sentences per MR combination in the training data, we used additional parameters to create a PERSONAGE "personality", which uses pragmatic markers such as "rather", "somewhat", and "actually". We prioritized sentences *without* these additional pragmatic markers when creating our dataset.

We also experiment with repeating instances in the training data to further expand our dataset. We repeat each instance three times, thus tripling the size of our training data from around 40k instances to around 120k. This method, as well as a variation of it where we compute permutations of the MR to supplement our training data, is discussed further in our companion submission to the challenge, "TNT-NLG, System 2".

4 Model Description

For our seq2seq model, we build on sequence to sequence TGEN (Dušek and Jurcicek, 2016b). It utilizes a sequence of LSTMs (Hochreiter and Schmidhuber, 1997) for the encoder and the decoder. We utilize an additional context encoder, which is an RNN based encoder (Dušek and Jurcicek, 2016a), and takes in PERSONAGE generated output for the same input MR as a prior input. The decoder has an attention mechanism (Bahdanau et al., 2015) which generates a weighted average of the encoder states, including the input and the context encoder states at all time steps. We use beam search during decoding with a beam size of 10.

The beam outputs generated by the sequence neural model is channeled to the reranking classifier with an additional encoder framework and a classification layer that identifies the different slot names and dialogue acts that have been realized in the output. The reranking classifier then outputs a binary vector indicating the presence or absence of dialogue act tags, attribute name and values in the generated sentence. After decoding, the input MR is also converted to a similar binary vector and the reranking penalty is the weighted Hamming distance between the classification output of the reranking classifier with the input MR sequence. Next, the sentences with the highest scores are chosen and passed to a ngram ranker which maximizes the BLEU score between

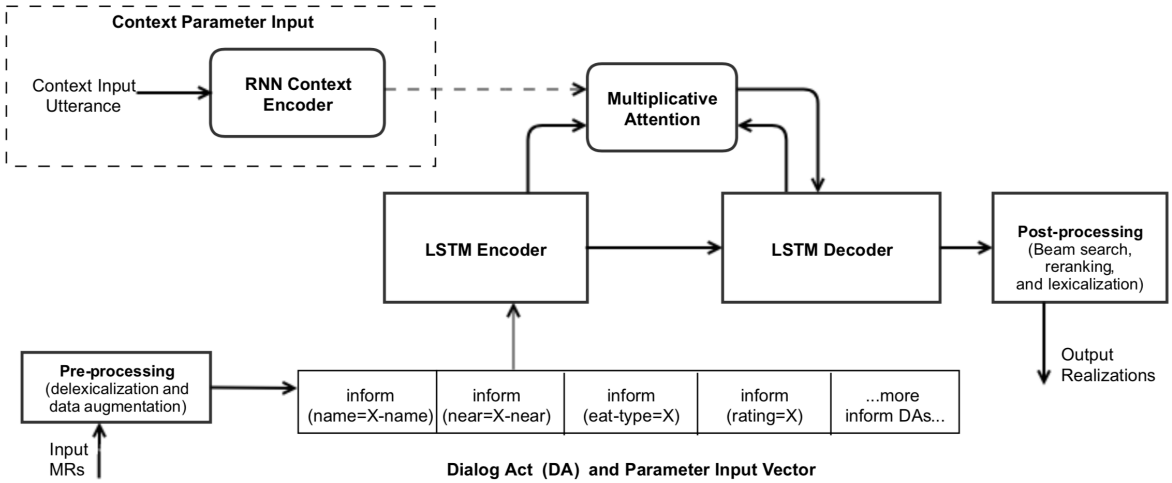

Figure 1: System Architecture

the outputs and the context input. The sentence with the highest score is ultimately picked as the output sentence. Figure 1 shows the structure of our main model.

System 1, Primary System: The model described above in Section 4, with input MRs from the E2E challenge and with PERSONAGE generated sentences as context.

System 1, Model 1: This is an additional augmentation technique we perform on the above model where we increase the size of our data by three times by repeating the input E2E MRs.

4.1 Training Setup

We used the 42k training samples from the E2E challenge for both model setups described in Section 4. We experimented with TGen default parameter settings as well as other variations, finally training for 20 epochs with 0.001 as the learning rate for the reranking classifier and 0.0005 for the decoder, with teacher forcing and 4000 validation sentences. We used batches of 20 and cross entropy as the loss function for training both the classifiers in all three setups. For Model 2, we trained on 120k samples, which was the direct result of the repeating of training data we performed.

We generated samples for the 630 MRs in the E2E Challenge test set for the experiments described in Section 5. During decoding, we used a beam search with a beam size of 10, making use of the ranking classifier to rank each of the beam outputs.

5 Results

In this section, we present model outputs and show evaluations of our system from the E2E Challenge, using automated metrics and qualitative human evaluation from CrowdFlower (Dušek et al., 2018).

5.1 Sample Output

Table 2 shows system output for a single meaning representation (MR) from each of our models. In Row 1, we see an example of a simple MR with only 4 slots (NAME, EAT-TYPE, CUSTOMER-RATING, and NEAR). For both of our model outputs, we see that the slots are realized in a single sentence. While the primary system uses "with" as an aggregation operator to combine the distinct propositions, Model 1 demonstrates a more basic sentence realization.

Row 2 shows an example of a more complex MR with 7 slots. In this case, our primary system realizes the output in two distinct sentences, grouping NAME, FAMILY-FRIENDLY, EAT-TYPE, FOOD, AND NEAR in the first sentence, and choosing to leave the PRICE-RANGE slot for a single, simple secondary sentence. We also see that the primary system learns to express the PRICE-RANGE as "high price range" from other training examples from the E2E data. Model 1 aggregates all slots into a single sentence.

5.2 Automatic Metrics

The E2E challenge evaluation included the use of automated metrics for judging system output and ranking competing systems. Here, we present the

#	MR	System	Sample Output
1	NAME[COCUM] EATTYPE[COFFEE SHOP] CUSTOMER RATING[LOW] NEAR[EXPRESS BY HOLIDAY INN]	Primary	With a low customer rating, Cocum is a coffee shop near Express by Holiday Inn.
		Model1	Cocum is a low rated coffee shop near Express by Holiday Inn.
2	NAME[THE WRESTLERS] EATTYPE[RESTAURANT] FOOD[JAPANESE] PRICERANGE[MORE THAN £30] AREA[RIVERSIDE] FAMILYFRIENDLY[YES] NEAR[RAJA INDIAN CUISINE]	Primary	The Wrestlers is a children friendly japanese restaurant in riverside near Raja Indian Cuisine. It has a high price range.
		Model1	The Wrestlers is a children friendly japanese restaurant in the riverside area near Raja Indian Cuisine with a price range of more than £30.

Table 2: System output for the same MRs across each model

results of the automatic evaluation using the script provided by the E2E NLG Challenge Evaluation Metrics[2]. The script calculates scores for the following metrics:

- BLEU: n-gram precision
- NIST: weighted n-gram precision
- METEOR: n-gram with synonym recall
- ROUGE: n-gram recall
- CIDEr: weighted n-gram cosine similarity

Table 3 summarizes the results of the automated metrics for our models on the test set of 630 MRs. We include the TGEN baseline results for reference from the TGen system by Dusek et al. (2016b), which is a seq2seq model with attention (Bahdanau et al., 2015) and beam search, including a reranker to penalize outputs to prevent straying from the input MR. Our results show that our primary system and Model 1 score comparably to the baseline for all metrics, and both score higher for METEOR.

We also observe that our primary system scores higher than Model 1 for all metrics, noting that the primary system uses PERSONAGE for context, but does not repeat training instances. Our best METEOR score of 0.4517 comes from the primary model, which scores the 5^{th} highest METEOR score in the competition, where the highest score is 0.4571. This is likely due to the stylistic variability of our outputs as compared to the reference texts for E2E, due to the variation introduced by Personage, which we plan to explore in more detail for future work.

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

5.3 Human Evaluation

Human evaluation for the competition was conducted by the organizers, using CrowdFlower, on the 19 primary systems and the baseline model from Dusek et al. (2016b), for a total of 20 systems. The evaluation involved showing crowd workers five randomly selected system outputs and a matching human NL output for reference, and asking them to rank the outputs from best to worst (allowing ties).

The rankings were based on two separate metrics, *quality* and *naturalness*. Quality takes into account overall quality metrics, such as grammatical correctness, fluency, adequacy, which would be considered the primary metrics for direct application in a real NLG system. Crowd workers were presented the MR along with the system outputs when judging for quality. Naturalness considers how likely it is that the output could have been produced by a native speaker. The MRs were not shown to crowd workers when judging for naturalness.

The competition evaluation results were then computed using the TrueSkill algorithm from Sakaguchi et al. (2014). For quality, 1,260 pairwise comparisons per system, or 25,200 comparisons in total, were made. For naturalness, 1,890 pairwise comparisons were made (37,800 total comparisons). Systems were then ranked by their TrueSkill scores, and then clustered, such that different systems within the same cluster are considered tied. Clustering was done using bootstrap resampling ($p \leq 0.05$)[3].

[3]More detail about the evaluations can be found on the challenge homepage: http://www.macs.hw.ac.uk/InteractionLab/E2E/

Model	BLEU	NIST	METEOR	ROUGE_L	CIDEr
Baseline	0.6593	8.6094	0.4483	0.6850	2.2338
Sys1-Primary	**0.6561**	**8.5105**	**0.4517***	**0.6839**	**2.2183**
Sys1-Model1	0.6476	8.4301	0.4508*	0.6795	2.1233

Table 3: Test Evaluation Results: Automatic Metrics (Our best scores are in bold. An * indicates scores higher than the baseline.)

Quality: For our quality evaluation, our primary system ranked in the second quality cluster. The first quality cluster consisted of a single system.

Naturalness: For our naturalness evaluation, our primary system ranked in the second quality cluster. Again, only one system scored within the first quality cluster.

6 Conclusions

In this paper, we present our first system submission to the E2E NLG challenge, showing how we massively augment the training data size to supply our neural generation models with more examples to learn from. We demonstrate how we utilize the PERSONAGE statistical language generator to greatly expand the input data size by generating synthetic data for training, and how we further augment the data by repeating instances in training. From our results, we find that we are able to achieve scores that are competitive with the baseline for all metrics, and higher than the baseline for METEOR. Also, we score competitively in the qualitative human annotation metrics.

Given the requirements of this task, the main goal was to generate simple, coherent realizations that closely mimic the trends in the data for input MRs, avoiding any variations in realization that might result in model penalization. However, we note that most of the work to date in neural NLG has focused on ensuring that the outputs faithfully realize the content specified in the input MRs. Earlier models for statistical natural language generation, such as the PERSONAGE engine we use for data augmentation, provide methods for training the statistical model to generate stylistic variations in responses. Thus, given that we have demonstrated an ability to retain some semantic fidelity in our realizations, we turn to the task of generating stylistically varied realizations for future work, focusing on the Big Five personality-based variations that PERSONAGE can produce.

Acknowledgements

We would like to thank the E2E NLG Challenge committee for organizing the competition and providing helpful review feedback.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *32nd International Conference on Machine Learning* .

Laurence Devillers, Hélène Maynard, Sophie Rosset, Patrick Paroubek, Kevin McTait, Djamel Mostefa, Khalid Choukri, Laurent Charnay, Caroline Bousquet, Nadine Vigouroux, et al. 2004. The french media/evalda project: the evaluation of the understanding capability of spoken language dialogue systems. In *LREC*.

Ondřej Dušek and Filip Jurcicek. 2016a. 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, pages 185–190. https://doi.org/10.18653/v1/W16-3622.

Ondřej Dušek and Filip Jurcicek. 2016b. 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)*. Association for Computational Linguistics, pages 45–51. https://doi.org/10.18653/v1/P16-2008.

Ondrej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge. In *(in prep.)*.

M. Gašic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, K. Yu, and S. Young. 2008. Training and evaluation of the his-pomdp dialogue system in noise. *Proc. Ninth SIGdial, Columbus, OH* .

Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. 2016. Contextual lstm (clstm) models for large scale nlp tasks. *KDD Workshop on Large-Scale Deep Learning for Data Mining* .

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

David Howcroft, Crystal Nakatsu, and Michael White. 2013. Enhancing the expression of contrast in the sparky restaurant corpus. In *Proceedings of the 14th European Workshop on Natural Language Generation*. Sofia, Bulgaria, pages 30–39.

Hiroaki Kitano, Minoru Asada, Itsuki Noda, and Hitoshi Matsubara. 1998. Robocup: Robot world cup. *IEEE Robotics & Automation Magazine* 5(3):30–36.

Gerasimos Lampouras and Andreas Vlachos. 2016. Imitation learning for language generation from unaligned data. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING*. ACL, pages 1101–1112.

Benoit Lavoie and Owen Rambow. 1997. A fast and portable realizer for text generation systems. In *Proc. of the Third Conference on Applied Natural Language Processing, ANLP97*. pages 265–268.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '10, pages 1552–1561. http://dl.acm.org/citation.cfm?id=1858681.1858838.

François Mairesse and Marilyn A Walker. 2008. A personality-based framework for utterance generation in dialogue applications. In *AAAI Spring Symposium: Emotion, Personality, and Social Behavior*. pages 80–87.

Francois Mairesse and Marilyn A. Walker. 2010. Towards personality-based user adaptation: Psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction* 20(3):227–278. https://doi.org/10.1007/s11257-010-9076-2.

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 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL HLT)*. San Diego, CA, pages 720–730.

Igor A Mel'cuk. 1988. Dependency syntax: Theory and practice. SUNY Press.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Saarbrücken, Germany.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing nlg data: Pictures elicit better data. In *International Conference on Natural Language Generation*.

Joseph Polifroni, Lynette Hirschman, Stephanie Seneff, and Victor Zue. 1992. Experiments in evaluating interactive spoken language systems. In *Proc. of the DARPA Speech and NL Workshop*. pages 28–33.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*. pages 1–11. http://aclweb.org/anthology/W/W14/W14-3301.pdf.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*. pages 3295–3301.

Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. 2016. Two are better than one: An ensemble of retrieval-and generation-based dialog systems. *arXiv preprint arXiv:1610.07149* .

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar, editors, *HLT-NAACL*. The Association for Computational Linguistics, pages 196–205. http://dblp.uni-trier.de/db/conf/naacl/naacl2015.html#SordoniGABJMNGD15.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialogue systems. In *Meeting of the Association for Computational Linguistics*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *Deep Learning Workshop at the 32nd International Conference on Machine Learning* .

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei hao Su, David Vandyke, and Steve J. Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *EMNLP*. pages 1711–1721.

Steve Whittaker, Marilyn Walker, and Johanna Moore. 2002. Fish or fowl: A Wizard of Oz evaluation of dialogue strategies in the restaurant domain. In *Language Resources and Evaluation Conference*.