TNT-NLG, System 2: Data Repetition and Meaning Representation Manipulation to Improve Neural Generation

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

Abstract

End-to-End (E2E) neural models that learn and generate natural language sentence realizations in one step have recently received a great deal of interest from the natural language generation (NLG) community. In this paper, we present "TNT-NLG" System 2, our second system submission in the E2E NLG challenge, which focuses on generating coherent natural language realizations from meaning representations (MRs) in the restaurant domain. We tackle the problem of improving the output of a neural generator based on the open-source baseline model from Dusek et al. (2016) by vastly expanding the training data size by repetition of instances in training, and permutation of the MR. We see that simple modifications allow for increases in performance by providing the generator with a much larger sample of data for learning. Our system is evaluated using quantitative metrics and qualitative human evaluation, and scores competitively in the challenge.

1 Introduction

Until recently, natural language generation (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). However, there has been a recent movement from the NLG community 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, 2016; 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), requiring that the generated utterance is similar to human-generated reference texts, and would be rated highly by humans in terms of quality and naturalness.

In this paper, we describe our second system submission to the E2E NLG challenge. We based our models on the open source baseline model from Dusek et al. (Dušek and Jurcicek, 2016), and focused our experimentation on how data augmentation through data repetition and MR manipulation results in a more robust model. We also experiment with the addition of synthetic data (using the PERSONAGE statistical generator) as noise can affect the quality of outputs produced[1]. We present one primary system, with three additional models based on input variations and parameter tuning. We describe our models and our methodology for vastly expanding the available training data, and show through competition evaluation that our models outperform the baseline for several evaluation metrics (specifically, BLEU, NIST, and ROUGE), and also score competitively on qualitative measures.

2 Related Work

Currently, most neural natural language generation (NNLG) models consist of a sequence to sequence (seq2seq) framework with multiplicative attention. In addition, beam search is often used at the decoding stage (as opposed to greedy decoding). The TGen model (Dušek and Jurcicek, 2016) trains a classifier to detect the presence of slots and values, and checks the number of deviations from the input MR.

Using context to inform NNLG representation and generate better realizations from input MRs is also highly popular. In dialogue generation, Sordoni et al. (2015) propose a simple approach incorporating the previous set of responses as a bag

[1] Using the PERSONAGE statistical generator for data augmentation is the main contribution of our companion E2E submission, "TNT-NLG System 1".

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 the weights for the input, and forget and output gates are learned. Our models in this work use synthetic data as a way to include extra context for training and improving the learning process for the generator.

The abundance of structured data in the restaurant domain makes it a good candidate for challenges like E2E, particularly because much of the previous work on NLG has focused on this domain (Howcroft et al., 2013; Polifroni et al., 1992; Whittaker et al., 2002; Stent et al., 2004; Devillers et al., 2004; Gašic et al., 2008; Mairesse et al., 2010). In this work, we find that we can improve the performance of NNLG even further by vastly expanding the amount of data available to the generator through data repetition and MR manipulation with minimal effort to generate such data, showing the potential portability of our approach to other domains in future work.

3 Data

E2E NLG datasets in the past were small and delexicalised, e.g BAGEL or RoboCup (Mairesse et al., 2010; Kitano et al., 1998). The data from this challenge is within the restaurant domain, where 50k instances have been crowdsourced and have an open vocabulary, complex syntactic structures and diverse discourse phenomena (Novikova et al., 2016).

The input data is in the form of meaning representation (MR), which includes dialogue attribute and value pairs. The MRs have on average 8.1 corresponding crowd-sourced natural language realizations (references) (Novikova et al., 2016). There are 8 different attributes, each with at least 2 values. 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 data is divided into three sets: train, dev and test. 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.

4 Model Description

All of our models are variants of a single system, trained with either different hyper-parameters, data, or for a different duration of time. Figure 1 shows the architecture of our main model.

For training our models, the input MR for slots such as the name of a restaurant and its locations can have infinite possibilities, since they are proper nouns. We delexicalize these slot values to tokens "X-name" and "X-near" for the name of the restaurant and the name of the restaurant that it's near. The model is trained on these delexicalized MRs paired with the training sentences. During post-processing, we lexicalize the outputs by replacing these placeholders with their proper nouns. The input vocabulary for the seq2seq model consists of different dialogue acts, the slot names, and the slot values. The output vocabulary consists of all the tokens found in the training labels.

For our seq2seq model, we build on the TGEN baseline system (Dušek and Jurcicek, 2016). It utilizes a sequence of LSTMs (Hochreiter and Schmidhuber, 1997) for the encoder and the decoder. The decoder has an attention mechanism (Bahdanau et al., 2015) which generates a weighted average of the encoder states at all time steps. We use beam search during decoding with a beam size of 10. The beam outputs generated by the neural model are 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 names 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 the outputs and the context input. The sentence with the highest score is ultimately

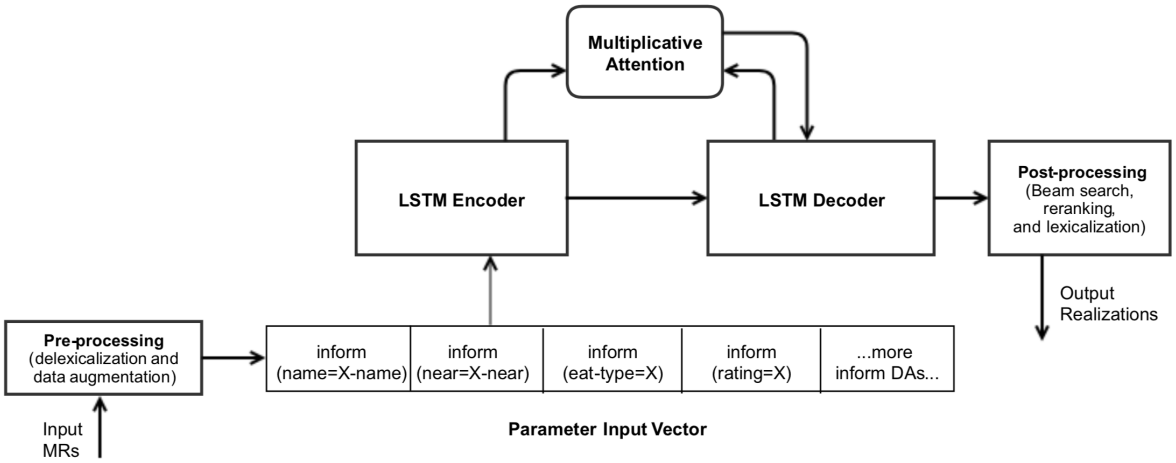

Figure 1: System Architecture

picked as the output sentence.

Since our input consists of meaning representations that do not have positional dependence on each other but are being fed to a RNN encoder, we tried data augmentation to fix the sparsity of our dataset. Our first form of data augmentation consists of simply expanding the training by repeating training instances. We triple the size of our data form 40k instances to 120k instances in this way.

Our second form of data augmentation focuses on meaning representation permutation. Experiments with the order of MRs have been performed by Nayak et al., (Nayak et al., 2017) showing significant improvements in the generated output. We permuted the MRs in the input and randomly chose three permutations, each one mapping to the same output as the original. The permutations were added to a new dataset that we ran our model on. Our use of three permutations guarantees that the model will see each word in its input vocabulary at least three times. Our training sample was initially about 40k. After incorporating new permutations, it grew to 120k. We describe our model variations below.

For each model in our system, we trained and tested on both our development and test sets under both data augmentation settings: data repetition and MR permutation. For the purposes of the E2E competition, we only submit our models with data repetition, but we include our MR permutation model results here for comparison, although they were generated post-competition.

System 2, Primary: For our primary system, we trained on the 120k samples generated from either the data repetition or meaning representation permutation datasets. We trained the seq2seq model and the reranking classifier for 50 epochs each.

System 2, Model 1: Model 2 is the same as Model 1, except that both the seq2seq and reranking classifiers have been trained for 20 epochs each.

System 2, Model 2: For this model, we trained 120k samples for 10 epochs and performed early stopping if the validation loss did not decrease for five consecutive iterations.

System 2, Model 3: We expanded our dataset to add data from a statistical tool PERSONAGE, an NLG engine that performs content planning, as noise to our dataset (Mairesse and Walker, 2008). PERSONAGE takes in the same input MRs and produces outputs with stylistic variations using aggregation operations and pragmatic markers. We randomly sample 30k sample outputs from the set generated by PERSONAGE to augment our training set, making it a total of 150k samples. More information on this data augmentation method is presented in a companion paper on our first E2E challenge submission, "TNT-NLG, System 1"[2].

[2]System1-Model1 in our companion paper is trained on the same data as System2-Model3 described here, but uses context encoding.

4.1 Training Setup

We used the 40k training samples from the E2E challenge for all the model setups described in Section 4. We experimented TGen default parameter settings as well as other variations, finally training them 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 classifiers in all three setups. For Model 3, we trained the model on 150k training samples which included stylistically varied output generated from Personage to augment our data in terms of style and variation.

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 our results and evaluation from the E2E competition. First, we show output from each of our systems, and present the competition evaluation of each model using both automated metrics and qualitative human evaluation from CrowdFlower (Dušek et al., 2018). Again, we point out that only our "Expanded Training with Data Repetition" model was an official competition submission, but we include our MR permutation results for comparison.

5.1 System Output

Realizations for a single meaning representation (MR) for each model in our data repetition system are shown in Table 1. For both MRs, we see a highly diverse set of realizations across the models.

In Row 1, Model 3 is the only model that aggregates all slots into a single sentence, whereas Model 2 realizes the content into 3 sentences. We see strong variation in how the propositions are presented, ranging from a simple *"It is near Cafe Rouge."* with Model 2, to much more complex aggregations, such as *"Near Cafe Rouge in the city centre is a high priced English restaurant called the Cricketers."*. The stylistic variability of our outputs as compared to the reference texts for E2E is due to the variation introduced by PERSONAGE, which we plan to explore in more detail for future work.

We see another example of single-sentence aggregation in Row 2, Model 3, where again all slots (8 in this example) are aggregated into a single, complex realization. The other model realizations consist of at least 2 sentences, with around 3 slots per sentence.

5.2 Automatic Metrics

Quantitative evaluation of our systems is based on the automatic metrics used for judging system output and ranking competing systems from the E2E Challenge. The E2E NLG Challenge Evaluation Metrics[3] 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

We present the automatic evaluation scores for each of our models for dev (547 MRs) and test (630 MRs) in Table 2 and Table 3, respectively, and also include the TGEN baseline results for reference from the TGen system by Dusek et al. (2016). The baseline model is a seq2seq model with attention (Bahdanau et al., 2015) and beam search and reranking for penalization. We present both our official submission the E2E Challenge ("Expanded Training with Data Repetition"), and our post-competition results for "Expanded Training with MR Permutation".

For our data repetition system, we observe from the results that our primary system and Models 1 and 2 all perform comparably to the baseline on the test and dev results, but Model 3 falls behind for all metrics. For the dev results, the primary system and Model 1 out performed the baseline for one metric each, NIST for the primary system, and ROUGE_L for Model 1. We also note that we do not have a single winning model across all metrics. Instead, we see, for the test results, that Model 1 scores best for BLEU (0.6606), NIST (6.622), and CIDEr (2.1997), beating the baseline and also achieving the 3^{rd} highest NIST score in the competition (the highest NIST score was 8.777). Model 2 scores best for METEOR (0.4482). Our best ROUGE_L score of 0.6853

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

#	MR	System	Sample Output
1	NAME[THE CRICKETERS] EATTYPE[RESTAURANT] FOOD[ENGLISH] PRICERANGE[HIGH] CUSTOMER RATING[AVERAGE] AREA[CITY CENTRE] FAMILYFRIENDLY[YES] NEAR[CAFE ROUGE]	Primary	The Cricketers is an english restaurant near Caf Rouge in the city centre. It is children friendly and has a high price range and an average customer rating.
		Model1	Near Cafe Rouge in the city centre is a high priced english restaurant called The Cricketers. It has an average customer rating and is children friendly.
		Model2	The Cricketers is a restaurant providing english food in the high price range. It is located in the city centre. It is near Cafe Rouge. Its customer rating is average and it is children friendly.
		Model3	The Cricketers is in city centre near Cafe Rouge with an average rating, it is an english restaurant and family friendly and it is expensive.
2	NAME[THE MILL] EATTYPE[PUB] FOOD[FAST FOOD] PRICERANGE[£20-25] CUSTOMER RATING[HIGH] AREA[RIVERSIDE] FAMILYFRIENDLY[YES] NEAR[CAFE ROUGE]	Primary	The Mill is a pub providing fast food in the £20-25 price range. It is located in the riverside area near Cafe Rouge. It has a high customer rating and is kid friendly.
		Model1	The Mill is a fast food pub near Cafe Rouge in the riverside area. It has a high customer rating and is kid friendly and has a price range of £20-25.
		Model2	The Mill is a pub providing fast food in the £20-25 price range. It is located in the riverside area near the Cafe Rouge. It has a high customer rating and is child friendly.
		Model3	The Mill is a fast food restaurant and kid friendly with a price range of £20-25 with a high rating, it is in riverside near Cafe Rouge and it is a pub.

Table 1: Data Repetition System output for the same MRs across each model

comes from the primary system, also beating the baseline.

For our post-competition MR permutation system on dev, we find that for all metrics other than NIST, our best results improve on our best results for our data repetition system (as well as the baseline). For test, the permutation system improves on all metrics other than NIST and ROUGE_L, and the results improve on the baseline for BLEU, NIST, and METEOR. These results follow the findings of Nayak et al. (2017) who also say significant improvements in the generated output when allowing for MR permutation. In future work, we plan to perform additional experiments where we both repeat the training data with and without MR permutation in an effort to further improve our output.

5.3 Human Evaluation

Qualitative evaluation of competing systems was done on CrowdFlower by the organizers, evaluating the 19 primary systems and Dusek et al.'s baseline system (2016). Crowd workers were shown five randomly selected system outputs for a single MR, and asked to rank them from best to worst, with ties allowed. The crowd workers were also shown a matching human NL output for comparison.

The metrics used for evaluation were *quality*, which is particularly important for real NLG systems, takes into account aspects such as grammatical correctness, fluency and adequacy, and *naturalness*, which considers how likely it would be that the output could have been produced by a native speaker. When judging for quality, crowd workers were shown an MR with the system out-

Model	BLEU	NIST	METEOR	ROUGE_L	CIDEr
Baseline	0.6925	8.4781	0.4703	0.7257	2.3987
Expanded Training with Data Repetition					
Sys2-Primary	**0.6921**	**8.6107***	0.4527	0.7188	2.2408
Sys2-Model1	0.6914	8.4047	**0.4650**	**0.7269***	**2.3441**
Sys2-Model2	0.6728	8.4708	0.4523	0.7068	2.2129
Sys2-Model3	0.3714	6.5365	0.3766	0.5170	1.6090
Expanded Training with MR Permutation					
Sys2-Primary-Permute	0.6709	8.4815*	0.4589	0.7066	2.2307
Sys2-Model1-Permute	**0.7195***	**8.6291***	**0.4686**	**0.7374***	**2.4548***
Sys2-Model2-Permute	0.6886	8.4842*	0.4654	0.7101	2.2757
Sys2-Model3-Permute	0.3762	6.2973	0.3658	0.5100	1.0943

Table 2: Dev Evaluation Results: Automatic Metrics (Our best scores are in bold. An * indicates scores higher than the baseline.)

Model	BLEU	NIST	METEOR	ROUGE_L	CIDEr
Baseline	0.6593	8.6094	0.4483	0.6850	2.2338
Expanded Training with Data Repetition					
Sys2-Primary	0.6502	8.5211	0.4396	**0.6853***	2.1670
Sys2-Model1	**0.6606***	**8.6223***	0.4439	0.6772	**2.1997**
Sys2-Model2	0.6563	8.5482	**0.4482**	0.6835	2.1953
Sys2-Model3	0.3681	6.6004	0.3846	0.5259	1.5205
Expanded Training with MR Permutation					
Sys2-Primary-Permute	0.6482	8.5844	0.4398	0.6801	2.1269
Sys2-Model1-Permute	0.6588	**8.6605***	0.4411	0.6801	2.1937
Sys2-Model2-Permute	**0.6619***	8.5653	**0.4502***	**0.6841**	**2.2151**
Sys2-Model3-Permute	0.2023	4.2048	0.2281	0.3846	0.5333

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

puts, but they were not shown one when judging for naturalness.

The TrueSkill algorithm was used to score the systems based on the annotations (Sakaguchi et al., 2014). Quality judgments involved 1,260 pairwise comparisons per system (25,200 comparisons in total), whereas naturalness involved 1,890 pairwise comparisons (37,800 in total). The TrueSkill scores were then used to rank and cluster the systems (using bootstrap resampling with $p \leq 0.05$).[4]

Here, we discuss our quality evaluations for the "Expanded Training with Data Repetition" model, our official submission to the E2E Challenge.

Quality: In terms of quality, our primary system was ranked in the second quality cluster. The first quality cluster consisted of a single system.

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

6 Conclusions

In this paper, we describe our second submission to the E2E NLG challenge, based on data

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

repetition and meaning representation manipulation to massively augment the size of our training data. From our quantitative results, we see an improvement over the baseline for BLEU, NIST, and ROUGE_L, and show that we score competitively against other teams in the qualitative evaluation. We see from our varied output that our models are able to learn to produce varied realizations, learning complex aggregation operations and outputting coherent and semantically sound output. For future work, we plan to explore better parameter tuning for our models and further analysis of the learning process based on the size of the training data available to the generator.

Acknowledgments

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. 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)*. 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.

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.

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.

Neha Nayak, Dilek Hakkani-Tur, Marilyn Walker, and Larry Heck. 2017. To plan or not to plan? sequence to sequence generation for language generation in dialogue systems.

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.

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*.

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*.