E2E NLG Challenge Submission: Towards Controllable Generation of Diverse Natural Language

Henry Elder
Adapt Centre / DCU
henry.elder@adaptcentre.ie

Sebastian Gehrmann
Harvard NLP
gehrmann@g.harvard.edu

Alexander O'Connor
Pivotus Ventures Inc.
alexander@pivotusventures.com

Qun Liu
Adapt Centre / DCU
qun.liu@dcu.ie

Abstract

We report experiments with natural language generation models that can be used in task oriented dialogue systems. We explore the use of additional input to the model to encourage diversity and control of outputs. While our submission does not rank highly using automated metrics, qualitative investigation of generated utterances points to interesting research directions.

1 Introduction

A dialogue act (DA) based meaning representation (MR) is a high level abstraction of information to be contained within a sentence. Natural language generation (NLG) from an MR requires a model to make low level decisions about syntax and sentence structure while accurately including the required knowledge from the DAs. There has been a trend in NLG towards the development of data driven models through the use of unaligned datasets (Mei et al., 2016; Wen et al., 2015).

The end-to-end (E2E) challenge (Novikova et al., 2017) focuses on using an MR to generate restaurant descriptions. The E2E dataset (Novikova et al., 2016) contains a wide vocabulary and complex sentence structures. As noted by Sharma et al. (2017) this is an improvement on previous datasets which were smaller and focused on less challenging NLG tasks. The E2E challenge requires participants to develop a natural language generator that can accurately verbalize the MR and use language in a way that is highly rated by humans.

Our approach is based on sequence-to-sequence neural machine translation models (Sutskever et al., 2014) which provide a strong baseline for correctly verbalizing MRs. In place of delexicalization of MRs we opt instead to use a pointer network (Vinyals et al., 2015), which allows the model to copy tokens directly from the source sequence into the generated utterance. We tackle the subsequent challenge of generating sentences that are interesting for humans by including an additional DA in the MR which influences the structure and diversity of generated utterances.

Table 1 demonstrates how our model includes an extra DA that allows for more direct control over words which appear in the generated utterance.

Meaning Representation
name[The Wrestlers]
eatType[restaurant]
food[Japanese]
priceRange[more than £30]
area[riverside]
familyFriendly[no]
near[Raja Indian Cuisine]
<u>additionalWords</u>[*looking adults offerings really try good prices situated*]
Generated utterance
If you're *looking* for an *adults* only Japanese restaurant, *try* The Wrestlers. It is *really good* and *situated* near Raja Indian Cuisine. The *prices* are more than £30.

Table 1: Utterance generated with a novel DA containing additional words

2 System Description

As the goal of the task is to maximize human ratings, we focus on increasing the diversity of the outputs. The E2E challenge could be viewed as a task similar to that of a restaurant proprietor or public reviewer creating a website description. For this reason we add an element of control to the model by including a customisable DA.

Typical approaches to generating diverse outputs focus on objective functions that affect the decoding step (Li et al., 2015). Our approach of augmenting the source sequence takes inspiration from recent work in paraphrase generation (Guu et al., 2017) and generating structured queries from natural language (Zhong et al., 2017). And is similar to previous work on common sense dialogue models (Young et al., 2017) and content-introducing text generation (Mou et al., 2016). Other approaches to controllable text generation have focused on more abstract inputs. Language models which generate text about a specific topic, product, person, sentiment (Li et al., 2016; Tang et al., 2016; Fan et al., 2017; Dong et al., 2017).

2.1 Additional words model

We augment the MR with an extra DA containing additional words to be included in the generated sentence. To obtain the data for this we looked at each target sentence and, using a set of rules, determined what words the model would learn to include. These selected words were added to the source sequence inside a custom DA. This ability of the model to accept additional words ensured that we would have both diversity of outputs and fine grained control over those outputs at test time.

For our additional words model we extracted tokens from the target sequence that adhered to the following set of rules:

- Not part of a list of stopwords
- Does not appear in the source sequence or meaning representation
- Does not contain punctuation or numbers

After the original list was compiled we removed the most frequently appearing token *located* and any tokens which occurred less than 6 times.

Table 2 contains an example of the training data used for the model.

2.1.1 Simulating choice of additional words

Table 3 shows how we simulate user choice of additional words by training a sequence-to-sequence model on a processed version of the training data.

The unique contents of each DA in the MR is treated as a single token. We omit the *name* and *near* DAs as they were observed to have little correlation with the semantics of the additional words chosen. The model attempts to correlate specific

Source sequence
name[The Vaults]
eatType[pub]
priceRange[more than £30]
customer rating[5 out of 5]
near[Café Adriatic]
additionalWords[*star Prices start*]
Target sequence
The Vaults pub near Café Adriatic has a 5 *star* rating. *Prices start* at £30.

Table 2: Example from the additional words model training set

DA with the additional words that appear in the target sentences.

Additional words are sampled from the model. We scale the final output layer of the model before applying softmax and sampling tokens for the generated utterance. The value used for scaling is known as *temperature*. Higher values of temperature lead to more diverse outputs. Temperature values close to 0 lead to the model choosing more conservative outputs. We use values of 0.9 to 1.1, to encourage the generation of a more diverse set of additional words.

Source sequence
pub
more_than_£30
5_out_of_5
Target sequence
star Prices start

Table 3: Example pair used for training the additional word generator

3 Experiments

The data set was tokenized using the NLTK port of the moses tokenizer with aggressive hyphen splitting. For each DA a custom start and stop token was added to the source sequence. [1]

The models used were from the OpenNMT-py library (Klein et al., 2017). Our model architecture contains 2 layers of bidirectional recurrent neural networks (RNN) with long short-term memory (LSTM) cells (Hochreiter and Schmidhuber, 1997). We use 500 hidden units for the encoder

[1] e.g. __name_start__ The Vaults __name_end__

Model	BLEU	NIST	METEOR	ROUGE-L	CIDEr
Additional words - temperature 1.1	0.5307	7.1738	0.4108	0.6112	1.5658
Additional words - temperature 1.0	0.5574	7.4078	0.4171	0.6308	1.6380
Additional words - temperature 0.9	0.5659	7.5196	0.4209	0.6327	1.7652
Baseline	0.6925	8.4781	0.4703	0.7257	**2.3987**
Additional words - extracted from target	**0.7381**	**9.9435**	**0.4726**	**0.7508**	2.2858

Table 4: Dev set results

Model	BLEU	NIST	METEOR	ROUGE-L	CIDEr
Additional words - temperature 1.1	0.5092	7.1954	0.4025	0.5872	1.5039
Additional words - temperature 1.0	0.5265	7.3991	0.4095	0.5992	1.6488
Additional words - temperature 0.9	0.5573	7.7013	0.4154	0.6130	1.8110
Baseline	**0.6593**	**8.6094**	**0.4483**	**0.6850**	**2.2338**

Table 5: Test set results

and decoder layer, and 500 units for the word vectors which are learned jointly across the whole model. We add dropout of 0.3 applied between the LSTM stacks.

The models are trained using Adam (Kingma and Ba, 2014) with learning rate 0.001 and learning rate decay of 0.5 applied after 8 epochs. The models were trained for 10 epochs and the best performing checkpoint on the development set was chosen.

The exploration and choice of hyperparameters was aided by the use of Bayesian hyperparameter optimization platform SigOpt (SigOpt, 2014).

Model	Naturalness	Quality
Baseline	2nd	2nd
Additional words - temperature 1.1	4th	4th

Table 6: True skill clusters

4 Results & Discussion

Table 4 shows evaluation results on the development of the baseline model (Novikova et al., 2017) and the additional words model with additional words generated with various temperatures. We also include the results of the additional word model with words extracted from a random target sentence in the corresponding multiple reference set. These results are consistent with the test set results in table 5. Table 6 contains rankings in the different true skill clusters for naturalness and quality (Dušek et al., 2018).

Automated evaluation and subsequent human evaluation results show our additional words model performs poorly relative to the baseline. A manual observation of the model's outputs reveal many errors such as repeated phrases and occasionally absent or incorrect information. We include a collection of generated utterances from the test set in table 7 to highlight areas where the model performs both well and poorly relative to the baseline.

Utterances from the baseline model tend to be more consistent but when viewed over many hundreds of samples this can be dry and repetitive. In most cases the baseline model appears to have learned its own simple templates for generating utterances from an MR. The template has a rough form that changes naturally depending on which DAs the model is required to include.

> *[name]* is a *[food] [eatType]* near *[near]* in the *[area]*. It has a *[customer rating]* and a price range of *[price range]*. It is *[family friendly]*.

Many verbalisation issues in the additional word model arise due to a conflict between an additional word and the existing DAs in the MR. The model used for generating additional words could be improved substantially. Increasing the minimum frequency of occurrence for additional words in the training data may give the model more examples from which to better learn correct syntax. The additional words model also suffers from an issue, common with pointer networks, in which source tokens are incorrectly repeated in the generated utterance. One way to handle this would be to have a second stage of training with a coverage loss as in See and Manning (2017).

MR Additional words DA	name[The Cricketers] eatType[coffee shop] customerrating[low] familyFriendly[no] near[Express by Holiday Inn] additionalWords[satisfaction considered Join]
Additional words model Baseline	The Cricketers is a coffee shop near Express by Holiday Inn. It has a low customer satisfaction rating and is not considered family-friendly. Join Us. The Cricketers is a coffee shop near Express by Holiday Inn. It has a low customer rating and is not family-friendly.
MR Additional words DA	name[Zizzi] eatType[coffee shop] area[riverside] additionalWords[Located along river side offering snacks]
Additional words model Baseline	Located along the river side , Zizzi is a coffee shop offering snacks . Zizzi is a coffee shop in the riverside area with a price range of less than 20.
MR Additional words DA	name[The Mill] eatType[restaurant] food[English] priceRange[high] customerrating[average] area[riverside] familyFriendly[yes] near[Café Rouge] additionalWords[offers fare higher bracket]
Additional words model Baseline	The Mill is a family friendly restaurant that offers English fare in the higher price bracket . It is located near Café Rouge in the riverside area . The Mill is an english restaurant near Café Rouge in the riverside area. It is children friendly and has a high price range and an average customer rating.
MR Additional words DA	name[Strada] eatType[coffee shop] customerrating[1 out of 5] near[All Bar One] additionalWords[try perfect however]
Additional words model **Baseline**	For a coffee shop near All Bar One, try Strada. It has a perfect customer rating of 1 out of 5 however . Strada is a coffee shop near All Bar One with a customer rating of 1 out of 5.
MR Additional words DA	name[Blue Spice] eatType[coffee shop] customerrating[average] near[Burger King] additionalWords[neat]
Additional words model **Baseline**	Blue Spice is a coffee shop located neat Burger King. Blue Spice is a coffee shop near Burger King. It has an average customer rating.
MR Additional words DA	name[Wildwood] eatType[pub] area[riverside] additionalWords[Located river serving]
Additional words model **Baseline**	Located near the river , Wildwood is a pub serving pub . Wildwood is a pub by the riverside.

Table 7: Example generated utterances using MRs from the test set

5 Conclusion

We proposed the use of an additional DA to improve the diversity and level of control over utterances. Results show both the underlying network and the method used for automatically generating additional words could be improved. Observation of high quality generated samples shows this to be an interesting research direction if such results can be obtained more consistently.

References

Emilie Colin, Claire Gardent, M Yassine, and Shashi Narayan. 2016. The WebNLG Challenge : Generating Text from DBPedia Data. *The 9th International Conference on Natural Language Generation*.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to Generate Product Reviews from Attributes.

Ondrej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge.

Ondej Dušek and Filip Jurčíček. 2016. Sequence-to-Sequence Generation for Spoken Dialogue via Deep Syntax Trees and Strings.

Angela Fan, David Grangier, and Michael Auli. 2017. Controllable Abstractive Summarization.

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):1–64.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2017. Generating Sentences by Editing Prototypes. 2.

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

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization.

Guillaume Klein, Yoon Kim, Yuntian Deng, Josep Crego, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source Toolkit for Neural Machine Translation. pages 67–72.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-Sequence Models for Parsing and Generation. pages 146–157.

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. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A Diversity-Promoting Objective Function for Neural Conversation Models.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. pages 994–1003.

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

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. What to talk about and how? Selective Generation using LSTMs with Coarse-to-Fine Alignment. *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16(*, pages 1–11.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to Backward and Forward Sequences: A Content-Introducing Approach to Generative Short-Text Conversation.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. ArXiv:1706.09254.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing NLG Data: Pictures Elicit Better Data. (1).

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.

Franois Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7-8):789–816.

Mark Przybocki, Kay Peterson, Sbastien Bronsart, and Gregory Sanders. 2009. The NIST 2008 Metrics for machine translation challenge—overview, methodology, metrics, and results. *Machine Translation*, 23(2):71–103.

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

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient Elicitation of Annotations for Human Evaluation of Machine Translation. *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT '14)*, pages 1–11.

Abigail See and Christopher D. Manning. 2017. Get To The Point : Summarization with Pointer-Generator Networks. *Association for Computational Linguistics*, pages 1–18.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of Unsupervised Metrics in Task-Oriented Dialogue for Evaluating Natural Language Generation.

Shikhar Sharma, Microsoft Maluuba, Jing He, Adeptmind Kaheer Suleman, Hannes Schulz, and Philip Bachman. 2016. Natural Language Generation in Dialogue using Lexicalized and Delexicalized Data.

Inc. SigOpt. 2014. Sigopt reference manual.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. pages 1–9.

Jian Tang, Yifan Yang, Sam Carton, Ming Zhang, and Qiaozhu Mei. 2016. Context-aware Natural Language Generation with Recurrent Neural Networks.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer Networks.

Bolin Wei, Shuai Lu, Lili Mou, Hao Zhou, Pascal Poupart, Ge Li, and Zhi Jin. 2017. Why Do Neural Dialog Systems Generate Short and Meaningless Replies? A Comparison between Dialog and Translation.

Tsung-Hsien Wen, Milica Gaši, Nikola Mrkši, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. pages 1711–1721.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A Network-based End-to-End Trainable Task-oriented Dialogue System. 1:438–449.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in Data-to-Document Generation.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2016. Topic Augmented Neural Response Generation with a Joint Attention Mechanism. pages 1–9.

Tom Young, Erik Cambria, Iti Chaturvedi, Minlie Huang, Hao Zhou, and Subham Biswas. 2017. Augmenting End-to-End Dialog Systems with Commonsense Knowledge.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning.