A surprisingly effective out-of-the-box char2char model on the *E2E NLG Challenge* dataset

Shubham Agarwal **Marc Dymetman**
NAVER Labs Europe*, Grenoble, France
shubhamagarwal92@gmail.com, marc.dymetman@naverlabs.com

Abstract

We train a char2char model on the *E2E NLG Challenge* data, by exploiting "out-of-the-box" the recently released *tf-seq2seq* framework, using some of the standard options of this tool. With minimal effort, and in particular without delexicalization, tokenization or lowercasing, the obtained raw predictions, according to a small scale human evaluation, are excellent on the linguistic side and quite reasonable on the adequacy side, the primary downside being the possible omissions of semantic material. However, in a significant number of cases (more than 70%), a perfect solution can be found in the top-20 predictions, indicating promising directions for solving the remaining issues.

1 Introduction

Very recently, researchers (Novikova et al., 2017) at Heriot-Watt University proposed the E2E NLG Challenge[1] and released a dataset consisting of 50K (MR, RF) pairs, MR being a slot-value Meaning Representation of a restaurant, RF (human ReFerence) being a natural language utterance rendering of that representation. The utterances were crowd-sourced based on pictorial representations of the MRs, with the intention of producing more natural and diverse utterances compared to the ones directly based on the original MRs (Novikova et al., 2016).

Most of the RNN-based approaches to Natural Language Generation (NLG) that we are aware of, starting with (Wen et al., 2015), generate the output word-by-word, and resort to special delexicalization or copy mechanisms (Gu et al., 2016) to handle rare or unknown words, for instance restaurant names or telephone numbers. One exception is (Goyal et al., 2016), who employed a char-based seq2seq model where the input MR is simply represented as a character sequence, and the output is also generated char-by-char; this approach avoids the rare word problem, as the character vocabulary is very small.

While (Goyal et al., 2016) used an additional finite-state mechanism to guide the production of well-formed (and input-motivated) character sequences, the performance of their basic char2char model was already quite good. We further explore how a recent out-of-the box seq2seq model would perform on E2E NLG Challenge, when used in a char-based mode. We choose attention-based *tf-seq2seq* framework provided by authors of (Britz et al., 2017) (which we detail in next section).

Using some standard options provided by this framework, and without any pre- or post-processing (not even tokenization or lowercasing), we obtained results on which we conducted a small-scale human evaluation on one hundred MRs, involving two evaluators. This evaluation, on the one hand, concentrated on the linguistic quality, and on the other hand, on the semantic adequacy of the produced utterances. On the linguistic side, vast majority of the predictions were surprisingly grammatically perfect, while still being rather diverse and natural. In particular, and contrary to the findings of (Goyal et al., 2016) (on a different dataset), our char-based model never produced non-words. On the adequacy side, we found that the only serious problem was the tendency (in about half of the evaluated cases) of the model to omit to render one (rarely two) slot(s); on the other end, it never hallucinated, and very rarely duplicated, material. To try and assess the potential value of a simple re-ranking technique (which we did not implement at this stage, but the

*Previously *Xerox Research Centre Europe*.

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

Proceedings of the SIGDIAL 2017 Conference, pages 158–163,
Saarbrücken, Germany, 15-17 August 2017.

approach of (Wen et al., 2015) and more recently the "inverted generation" technique of (Chisholm et al., 2017) could be used), we generated (using the beam-search option of the framework) 20-best utterances for each MR, which the evaluators scanned towards finding an "oracle", i.e. a generated utterance considered as perfect not only from the grammatical but also from the adequacy viewpoint. An oracle was found in the first position in around 50% of the case, otherwise among the 20 positions in around 20% of the cases, and not at all inside this list in the remaining 30% cases. On the basis of these experiments and evaluations we believe that there remains only a modest gap towards a very reasonable NLG seq2seq model for the E2E NLG dataset.

2 Model

Our model is a direct use of the seq2seq open-source software framework[2], built over TensorFlow (Abadi et al., 2016), and provided along with (Britz et al., 2017), with some standard configuration options that will be detailed in section 3. While in their large-scale NMT experiments (Britz et al., 2017) use word-based sequences, in our case we use character-based ones. This simply involves changing "delimiter" option in configuration files.

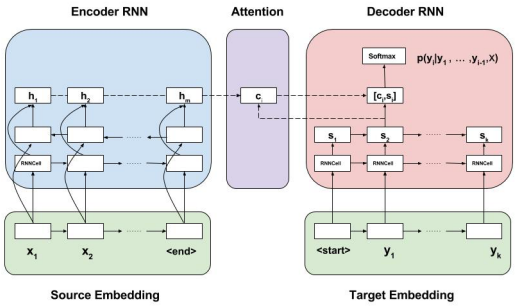

Figure 1: The seq2seq architecture of (Britz et al., 2017) (drawing borrowed from that paper). Contrary to word-based sequences, we use character-based sequences for generating grammatically correct and natural utterances.

Figure 1, borrowed from (Britz et al., 2017), provides an overview of the framework. While many options are configurable (number of layers, unidirectional *vs* bidirectional encoder, additive *vs* multiplicative attention mechanism, GRU (Cho et al., 2014) *vs* LSTM cells (Hochreiter and Schmidhuber, 1997), etc.), the core architecture is common to all models. This is by now a pretty standard attention-based encoder-decoder architecture based on (Bahdanau et al., 2015; Luong et al., 2015). The encoder RNN embeds each of the source words (in our case, characters) into vectors exploiting the hidden states computed by the RNN. The decoder RNN predicts the next word (resp. character) based on its current hidden state, previous character, and also based on the "context" vector c_i, which is an attention-based weighted average of the embeddings of the source words (resp. characters).

3 Experiments

3.1 Dataset

(Novikova et al., 2016) explain the protocol followed for crowdsourcing the *E2E NLG Challenge* dataset. Slightly different from the description in the article, there are two additional slots in the dataset: 'kidsFriendly' and 'children-friendly' which seem to be alternates for 'familyFriendly'. Thus, there are in total 10 slots (in decreasing order of frequency of being mentioned in the dataset MRs): name (100%), food (83%), customer rating (68%), priceRange (68%), area (60%), eatType (51%), near (50%), familyFriendly (25%), kidsFriendly (19%), children-friendly (19%). Also, the number of active slots in the MRs varies as: 3 (5%), 4 (17%), 5 (19%), 6 (19%), 7 (16%), 8 (4%).

3.2 Implementation

The *tf-seq2seq* toolkit (Britz et al., 2017) trains on pairs of sequences presented in parallel text format (separate source and target sequence files).[3] [4]

Taking cue from recommended configurations in Table 7 of (Britz et al., 2017) and the provided example configs in *tf-seq2seq*, we experimented with different numbers of layers in the encoder and decoder as well as different beam widths, while using the bi-directional encoder along with "additive" attention mechanism. As also observed

[2] https://github.com/google/seq2seq.

[3] We cleaned the E2E NLG Challenge data as there were a few erroneous newline characters (Line 603 in devset.csv as well as 30048 in trainset.csv). There were different character encodings for MR and RF, which we uniformized to utf-8. Also, there were a few wrongly encoded characters (such as on line 23191 in trainset.csv). We normalized these characters, after which there remained only two non-ascii characters: £ and é. Note: since submission, these issues have been corrected in the updated version of the Challenge data.

[4] Code for processing of the data, conversion to parallel text format as well as our configuration files for the tf-seq2seq model can be found at: https://github.com/shubhamagarwal92/sigdialSubmission/

Model Specification	Beam Width	Length Penalty	Depth (Number of layers)				
Encoder			1	1	2	4	4
Decoder			1	2	2	4	4
Cell Unit			GRU	GRU	GRU	GRU	LSTM
Greedy Search			20.94	22.59	23.5	23.84	23.98
Beam Search	Beam 5	0.0	15.85	22.47	21.76	22.73	20.15
	Beam 10	0.0	14.5	21.4	19.98	21.15	18.88
	Beam 20	0.0	13.48	20.18	18.5	19.94	17.93
	Beam 5	1.0	20.64	24.77	24.67	**24.94**	23.87
	Beam 10	1.0	21	25.05	24.88	24.69	24.27
	Beam 20	1.0	21.27	25.4	24.96	**24.6***	24.05

Table 1: BLEU scores on devset with different configuration: varying the depth of both encoder and decoder RNNs, type of cell unit, different beam width and length penalty. (Results reported for only a single experiment with training and prediction.)

by Britz et al. (2017), using a non-null "length-penalty" (alias length normalization (Wu et al., 2016)), significantly improved decoding results.

3.3 Results

We report the BLEU scores[5] for different configurations of the seq2seq model in Table 1. In our initial experiments, using a beam-width 5 (with no length penalty), with 4 layers in both the encoder and decoder and GRU cells, showed the best results in terms of BLEU (score of 24.94).

We observed significant improvements using length penalty 1, and decided to use this architecture as a basis for human evaluations, with a beam-width 20 to facilitate the observation of oracles. These evaluations were thus conducted on model [encoder 4 layers, decoder 4 layers, GRU cell, beam-width 20, length penalty 1] (starred in Table 1), though we found slightly better performing models in terms of BLEU at a later stage.

4 Evaluation

The human evaluations were performed by two annotators on the top 20 predictions of the previously discussed model, for the first 100 MRs of the devset, using the following metrics:

1. Semantic Adequacy
   a) Omission [1/0]: information present in the MR that is omitted in the predicted utterance (1=No omission, 0=Omission). **b) Addition [1/0]:** information in the predicted utterance that is absent in the MR (1=No addition, 0=Addition). **c) Repetition [1/0]:** repeated information in the predicted utterance (1=No repetition, 0=Repetition).
2. Linguistic Quality
   a) Grammar [1/0]: (1=Grammatically correct, 0=incorrect). Note: one annotator punished the model even for (rare) mistakes of punctuation. **b) Naturalness [2/1/0]:** subjective score to measure the naturalness of the utterance (2 being best). **c) Comparison to reference [1/0/-1]:** subjective score comparing the prediction with the crowdsourced RF. ('vsRef' in the Table 2, 1=Prediction better than RF, 0=Prediction at par with RF, -1=RF better than prediction).
3. Oracle [1/0/-1]: 1 if the first prediction is an "oracle" (i.e. considered as perfect, see section 1), 0 when the oracle is found in the top 20, and -1 when no oracle is found there.

5 Analysis

We show a few examples of utterances (predictions in first position, i.e. most probable) produced by our model, for discussion.[6]

1. **[MR]:** name[The Punter], customer rating[high], area[riverside], kidsFriendly[yes]
   [RF]: *In riverside area, there is The Punter, which is high rated by customers and kids are friendly.*
   [Pred]: The Punter is a kid friendly restaurant in the riverside area with a high customer rating.
2. **[MR]:** name[The Golden Palace], eatType[coffeee shop], food[Japanese], priceRange[£20-25], customer rating[high], area[riverside]
   [RF]: *For highly-rated Japanese food pop along to The Golden Palace coffee shop. Its located on the riverside. Expect to pay between 20-25 pounds per person.*
   [Pred]: The Golden Palace is a coffee shop providing Japanese food in the £20-25 price range. It is located in the riverside area.

[5] Calculated using multi-bleu perl script bundled with *tf-seq2seq*. Note that these results were computed on the original version of Challenge devset (updated recently) which did not group the references associated with the same MR, possibly resulting in lower scores than when exploiting multi-refs.

[6] Some more examples can be found in Table 4. The full list of human annotated examples, including the 20-best predictions and oracles, can be found at https://docs.google.com/spreadsheets/d/1wMu42g8bzyFxBUJ33QIdkqN3md3281pg6rLGrnDbEIE/edit?usp=sharing.

Ann	O(1/0)	A(1/0)	R(1/0)	G(1/0)	N(2/1/0)	vsRef(1/0/-1)	Or(1/0/-1)
Ann 1	51/49	100/0	97/3	93/7	85/13/2	46/16/38	50/18/32
Ann 2	51/49	100/0	98/2	98/2	80/18/2	29/36/35	51/18/31
Mean	51/49	100/0	97.5/2.5	95.5/4.5	82.5/15.5/2	37.5/26/36.5	50.5/18/31.5

Table 2: Human annotations for 100 samples using different metrics defined in Sec. 4. O (Omission), A (Addition), R (Repetition) and G (Grammar) are on binary scale. Naturalness is measured as (2/1/0) and Oracle as (1/0/-1). Predictions were also judged against the reference on a scale of (1/0/-1).

Slots	DA	Or@1	Or	No Or
3	1(1%)	1(100%)	0(0%)	0(0%)
4	29(29%)	24(83%)	3(10%)	2(7%)
5	25(25%)	13(48%)	6(24%)	6(28%)
6	29(29%)	11(34%)	5(17%)	13(48%)
7	11(11%)	1(9%)	3(27%)	7(64%)
8	5(5%)	1(20%)	1(20%)	3(60%)
Total	100	51	18	31

Table 3: Human annotations for different slots using beam-width 20. 'Or@1' represents the presence of an 'oracle' at first position while 'Or' represents the presence of 'Oracle' (desirable) in the top-20 predictions. Cases where no oracle was found are marked as 'No Or'.

3. **[MR]:** name[Strada], food[Fast food], priceRange [moderate], customer rating[1 out of 5], kidsFriendly[no], near [Rainbow Vegetarian Cafe]
[RF]: *Strada is a Fast food restaurant near the Rainbow Vegetarian caffe which has a moderate customer rating of 1 out of 5 for a non Kids friendly restaurant*
[Pred]: Strada is a moderately priced fast food restaurant in the **moderate price range**. It is located near Rainbow Vegetarian caffe.

Among the utterances produced by the model in first position (Pred), the most prominent issue was that of omissions (underlined in example 2). There were no additions or non-words (which was one of the primary concerns for (Goyal et al., 2016)). We observed only a couple of repetitions which were actually accompanied by omission of some slot(s) in the same utterance (repetition highlighted in bold in example 3). Surprisingly enough, we observed a similar issue of omissions in human references (target for our model). We then decided to perform comparisons against the human reference ('vsRef' in Table 2). Often, the predictions were found to be semantically or grammatically better than the human reference; for example observe the underlined portion of the reference in the first example. The two annotators independently found the predictions to be mostly grammatically correct as well as natural (to a slighty lesser extent).[7]

A general feeling of the annotators was that the predictions, while showing a significant amount of linguistic diversity and naturalness, had a tendency to respect grammatical constraints *better* than the references; the crowdsourcers tended to strive for creativity, sometimes not supported by evidence in the MR, and often with little concern for linguistic quality; it may be conjectured that the seq2seq model, by "averaging" over many linguistically diverse and sometimes incorrect training examples, was still able to learn what amounts to a reasonable linguistic model for its predictions.

We also investigate whether we could find an 'oracle' (perfect solution as defined in section 1) in the top-20 predictions and observed that in around 70% of our examples the oracle could be found in the top results (see Table 3), very often (51%) at the first position. In the rest 30% of the cases, even the top-20 predictions did not contain an oracle. We found that the presence of an oracle was dependent on the number of slots in the MR. When the number of slots was 7 or 8, the presence of an oracle in the top predictions decreased significantly to approximately 40%. In contrast, with 4 slots, our model predicted an oracle right at the first place for 83% of the cases.

6 Conclusion

We employed the open source *tf-seq2seq* framework for training a char2char model on the *E2E NLG Challenge* data. This could be done with minimal effort, without requiring delexicalization, lowercasing or even tokenization, by exploiting standard options provided with the framework.

Human annotators found the predictions to have great linguistic quality, somewhat to our surprise, but also confirming the observations in (Karpathy, 2015). On the adequacy side, omissions were the major drawback; no hallucinations were observed and only very few instances of repetition. We hope our results and annotations can help understand the dataset and issues better, while also being useful for researchers working on the challenge.

[7] Annotator-1 was more severe in highlighting even the (rare) punctuation issues as grammatical mistakes. There was also a slight disagreement with Annotator-2 being more severe than Annotator-1 when assessing the references against the predictions.

Slots	Type	Utterance
3	MR	name[Blue Spice], priceRange[£20-25], area[riverside]
	RF	*Blue Spice has items in the £20-25 price range and is in riverside.*
	Pred	Blue Spice is located in the riverside area with a price range of £20-25.
4	MR	name[The Punter], customer rating[high], area[riverside], kidsFriendly[yes]
	RF	*In riverside area, there is The Punter, which is high rated by customers and kids are friendly.*
	Pred	The Punter is a kid friendly restaurant in the riverside area with a high customer rating.
5	MR	name[Green Man], eatType[pub], food[English], area[city centre], near[Cafe Rouge]
	RF	*Green Man is a pub that can be found in the city centre, near caffe Rouge and serves English-style food.*
	Pred	Green Man is an English pub located in the city centre near caffe Rouge.
6	MR	name[The Golden Palace], eatType[coffeee shop], food[Japanese], priceRange[£20-25], **customer rating[high]**, area[riverside]
	RF	*For highly-rated Japanese food pop along to The Golden Palace coffee shop. Its located on the riverside. Expect to pay between 20-25 pounds per person.*
	Pred	The Golden Palace is a coffee shop providing Japanese food in the £20-25 price range. It is located in the riverside area.
7	MR	name[The Rice Boat], food[Chinese], priceRange[cheap], customer rating[average], area[city centre], **familyFriendly[no]**, near[Express by Holiday Inn]
	RF	*The Rice Boat is a not family friendly,cheap, average rated Chinese food restaurant near Express by Holiday Inn.*
	Pred	The Rice Boat provides Chinese food in the cheap price range. It is located in the city centre near Express by Holiday Inn. Its customer rating is average.
8	MR	name[The Eagle], eatType[coffeee shop], food[Japanese], priceRange[moderate], customer rating[1 out of 5], area[riverside], kidsFriendly[yes], near[Burger King]
	RF	*There is a one star mid priced family friendly coffee shop The Eagle near Burger King in the City centre. It offers Chinese food.*
	Pred	The Eagle is a kid friendly Japanese coffee shop in the riverside area near Burger King. It has a moderate price range and a customer rating of 1 out of 5.

Table 4: Sample predictions. For the first MR of each arity (3 to 8) in the devset, we show the best prediction of the model (the starred one in Table 1), along with the RF. Omissions of semantic material are highlighted in bold.

Acknowledgments We thank Éric Gaussier, Chunyang Xiao, and Matthias Gallé for useful suggestions.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR abs/1603.04467* .

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. ICLR*.

Denny Britz, Anna Goldie, Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. *CoRR abs/1703.03906* .

Andrew Chisholm, Will Radford, and Ben Hachey. 2017. Learning to generate one-sentence biographies from wikidata. *CoRR abs/1702.06235* .

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proc. EMNLP*.

Raghav Goyal, Marc Dymetman, and Eric Gaussier. 2016. Natural Language Generation through Character-based RNNs with Finite-State Prior Knowledge. In *Proc. COLING*. Osaka, Japan.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proc. ACL*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Andrej Karpathy. 2015. The unreasonable effectiveness of recurrent neural networks. http://karpathy.github.io/2015/05/21/rnn-effectiveness/.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2016. Character-based neural machine translation. In *Proc. ICLR*. pages 1–11.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proc. EMNLP*.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. 2017. The E2E NLG Shared Task .

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing NLG Data: Pictures Elicit Better Data. *CoRR abs/1608.00339* .

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proc. NIPS*. pages 3104–3112.

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

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klinger, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR abs/1609.08144* .