Noise and Neural Natural Language Generation: Rubbish in, Rubbish out?

Ondřej Dušek
Institute for Formal and Applied Linguistics
Charles University, Prague, Czech Republic
odusek@ufal.mff.cuni.cz

David M. Howcroft, Verena Rieser, & Karin Sevegnani
The Interaction Lab, MACS
Heriot-Watt University, Edinburgh
{D.Howcroft,V.Rieser,ks85}@hw.ac.uk

We examine the impact of noise on Neural Natural Language Generation (NNLG). First, we empirically determine how noisy training data impacts NNLG outputs. Second, how the type of noise in NNLG outputs (i.e. errors produced by NNLG) influence how humans perceive the quality of generated texts.

Neural Natural Language Generation (NNLG) is promising for generating text from Meaning Representations (MRs) in an "end-to-end" fashion, i.e. without the need for alignment (Wen et al., 2015, 2016a; Dušek and Jurčíček, 2016a; Mei et al., 2016). However, NNLG typically requires large volumes of in-domain data. This data is typically crowdsourced (e.g. Howcroft et al., 2017; Mairesse et al., 2010; Novikova et al., 2016; Wen et al., 2015, 2016b), which introduces noise. For example, in the E2E NLG Challenge, up to 40% of the data contains omitted or additional information (Dušek et al., 2019). In addition, NNLG can produce new types of errors, including ungrammatical outputs, misspellings from the training data, or even outputs which are semantically incorrect, i.e. hallucinating or omitting information (Gehrmann et al., 2018; Rohrbach et al., 2018).

Here we present the latest results in our ongoing investigation of the impact of several kinds of noise common in crowdsourced training datasets on NNLG for English:

- *Semantic noise*, which occurs when crowd workers write a text which differs from the source meaning representation they are tasked with communicating (i.e. inserting additional facts or dropping required information).
- *Typographic noise*, introduced when subjects spell a word incorrectly or make a typographic error. While this kind of noise may seem uninteresting, it is quite common in crowdsourced data and similar errors have been shown to have a dramatic impact on related neural models, e.g. in MT (Belinkov and Bisk, 2018).
- *Grammatical noise* is similar, resulting from disfluencies, non-standard syntax, and lack of punctuation.

Our results demonstrate the impact of semantic noise on two state-of-the-art neural generation models with different semantic control mechanisms, namely TGen (Dušek and Jurčíček, 2016a) and Semantically Controlled LSTMs (Wen et al., 2015). We use the E2E Challenge dataset (Dušek et al., 2019) as a noisy training set for these models and use a heuristic filtering script to produce a cleaned version of the dataset (E2E 1.1). This cleaned dataset contains a greater variety of MRs than the original corpus, since texts which were originally intended to be based off of the same MR are now differentiated when crowd workers in fact changed the intended meaning. We find that training on the clean data results in improvements for both systems with respect to automated word-overlap evaluation metrics like BLEU (Papineni et al., 2002) and reductions in semantic error rate for TGen.

The second aspect of our research is the impact of different kinds of errors in NNLG systems' *output* on the perceived system quality. To this end, we are developing a taxonomy of NNLG errors to understand what the most common kinds of problems are and conducting human evaluations of texts exhibiting varying degrees of these errors. The resulting human evaluation scores are further leveraged in training novel quality estimation models (as in Specia et al., 2010; Dušek et al., 2017). These models aim to predict the quality of an NLG output for an unseen MR without having access to any human-written reference texts, as is the case for word-overlap-based automatic metrics. This will allow their usage in production NLG systems for ranking candidate outputs and potentially improving overall output quality.

Acknowledgements

This research received funding from the EPSRC project MaDrIgAL (EP/N017536/1). Authors are listed alphabetically by last name.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation. In *ICLR*.

Ondřej Dušek and Filip Jurčíček. 2016a. 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*, pages 45–51, Berlin, Germany. arXiv:1606.05491.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2017. Referenceless Quality Estimation for Natural Language Generation. In *Proceedings of the 1st Workshop on Learning to Generate Natural Language (LGNL)*, Sydney, Australia. ArXiv: 1708.01759.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2019. Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Computer Speech & Language*.

Sebastian Gehrmann, Falcon Z. Dai, Henry Elder, and Alexander M. Rush. 2018. End-to-End Content and Plan Selection for Natural Language Generation. In *E2E NLG Challenge System Descriptions*.

David M. Howcroft, Dietrich Klakow, and Vera Demberg. 2017. The Extended SPaRKy Restaurant Corpus: Designing a Corpus with Variable Information Density. In *Proc. of Interspeech 2017*, pages 3757–3761, Stockholm, Sweden. ISCA.

Fran¸cois Mairesse, Milica Gasic, Filip Jurčícek, 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 (ACL)*, Uppsala, Sweden. Association for Computational Linguistics.

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 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, USA. arXiv:1509.00838.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing NLG data: Pictures elicit better data. In *Proceedings of the 9th International Natural Language Generation Conference*, pages 265–273, Edinburgh, UK. arXiv:1608.00339.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318. Association for Computational Linguistics.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24(1):39–50.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve J. Young. 2016a. A network-based end-to-end trainable task-oriented dialogue system. In *EMNLP*.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Lina Maria Rojas-Barahona, Pei-hao Su, David Vandyke, and Steve J. Young. 2016b. Multi-domain neural network language generation for spoken dialogue systems. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 120–129, San Diego, CA, USA. arXiv:1603.01232.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal.