Domain Adaptation for Natural Language Generation

Research Progress: First Year

Zdeněk Kasner

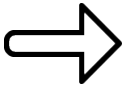
Supervisor: Ondřej Dušek

ÚFAL Seminar
24 September 2020

Topic Introduction

Natural Language Generation (NLG)

Data



Text

State of the Art in NLG

Template-based systems

- Pipeline of modules for subtasks
- Long-standing tradition
- Work in practice
- *Engineered for each domain*

Neural-based systems

- End-to-end, data-driven
- Inspired by recent advances in machine translation
- On-going research effort
- *Need data for each domain*

Domain Adaptation for NLG

Approaches

- Creating **synthetic data** for the target domain (Wen et al., 2016; Mi et al., 2019)
  - Works only for similar domains
- Using **domain-independent semantic representations** (Dethlefs, 2017; Tran and Nguyen, 2018; Tseng et al., 2019)
  - Restricts the input format
  - Requires additional annotations
- Finetuning the general-domain **pretrained models** (Chen et al., 2020)
  - Lack of control over the output

My Research

My Research

Research Focus

- **Task:** verbalize *all* and *only* the input data
- **Approach:** pretrained neural models
- **Problems:** lack of control, omissions and hallucinations

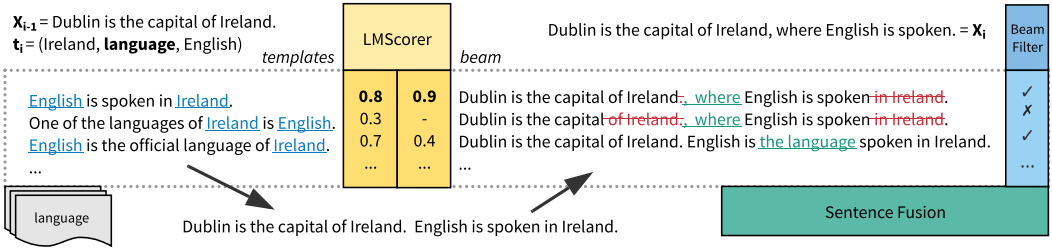
Solutions

1. Verbalize the data iteratively (→ *increase the control over the model)*
2. Limit the scope of the neural model (→ *prevent omissions and hallucinations)*

```
<entry>
   <tripleset><triple>Trane | foundingDate | 1913-01-01</triple>
   <triple>Trane | location | Ireland</triple>
   <triple>Trane | foundationPlace | La_Crosse,_Wisconsin</triple>
   <triple>Trane | numberOfEmployees | 29000</triple></tripleset>
   <lex>Trane, which was founded on January 1st 1913 in La Crosse, is based in Ireland.
       It has 29,000 employees.</lex>
</entry>
```

Approach

1. Select the best template for each data item using a language model
2. For each data item:
   - Fuse the filled template with the existing text
   - Check the text for consistency
3. Output the final text



Experiments

Ingredients

- **GPT-2** (Radford et al., 2019) – a pretrained language model
- **LaserTagger** (Malmi et al., 2019) – a text-editing model based on BERT

Datasets

- **WebNLG** (Gardent et al., 2017) – RDF triples from DBpedia + descriptions
- **Cleaned E2E** (Dušek et al., 2019) – restaurant attributes + descriptions

Results

- Substantial **improvements over the baseline** (although lacking behind SOTA)
- Can guarantee **zero entity errors** (at the cost of text fluency)
- **Zero-shot domain adaptation** with a sentence fusion dataset (Geva et al., 2019)

Under review for COLING 2020

Other Endeavors

- **Duolingo STAPLE Shared Task (WNGT 2020)**
  - Generating paraphrases for translations in 5 languages
  - Rather unsuccessful experiments with Levenshtein Transformer (Gu et al., 2019) for CUNI submission (Libovický et al., 2020)
- **WebNLG Challenge 2020**
  - Generating descriptions for DBpedia data in English and Russian
  - Good results with finetuning mBART (Liu et al., 2020)
  - Submission sent
- **Evaluating Semantic Accuracy of NLG**
  - Using a model finetuned for natural language inference to check the semantic accuracy of text
  - Co-authored a paper by O. Dušek, under review for INLG 2020

Future Research

Future Research

Follow-up

- Improving the text-editing approach
  - FELIX – text-editing model with arbitrary reordering (Mallinson et al., 2020)
  - Few-shot domain adaptation
- Wrapping up & publishing the code

Bigger Picture

- More diverse NLG datasets
  - Table-to-text
  - Commonsense reasoning
  - Logical reasoning
- Beyond the teacher accuracy $\rightarrow$ reinforcement learning?
- Online demo, visualizations

References

References I

Chen, Z., Eavani, H., Chen, W., Liu, Y., and Wang, W. Y. (2020). Few-shot NLG with pre-trained language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Online.

Dethlefs, N. (2017). Domain transfer for deep natural language generation from abstract meaning representations. *IEEE Computational Intelligence Magazine*, 12(3):18–28.

Dušek, O., Howcroft, D. M., and Rieser, V. (2019). Semantic Noise Matters for Neural Natural Language Generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426, Tokyo, Japan.

Gardent, C., Shimorina, A., Narayan, S., and Perez-Beltrachini, L. (2017). The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133.

Geva, M., Malmi, E., Szpektor, I., and Berant, J. (2019). DiscoFuse: A large-scale dataset for discourse-based sentence fusion. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3443–3455, Minneapolis, Minnesota.

Gu, J., Wang, C., and Zhao, J. (2019). Levenshtein transformer. In *Advances in Neural Information Processing Systems*, pages 11181–11191.

References II

Libovický, J., Kasner, Z., Helcl, J., and Dušek, O. (2020). Expand and filter: Cuni and lmu systems for the wngt 2020 duolingo shared task. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 153–160.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *arXiv:2001.08210 [cs]*. arXiv: 2001.08210.

Mallinson, J., Severyn, A., Malmi, E., and Garrido, G. (2020). Felix: Flexible text editing through tagging and insertion. *arXiv:2003.10687 [cs]*. arXiv: 2003.10687.

Malmi, E., Krause, S., Rothe, S., Mirylenka, D., and Severyn, A. (2019). Encode, tag, realize: High-precision text editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5057–5068.

Mi, F., Huang, M., Zhang, J., and Faltings, B. (2019). Meta-learning for low-resource natural language generation in task-oriented dialogue systems. *arXiv:1905.05644 [cs]*. arXiv: 1905.05644.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Tran, V.-K. and Nguyen, L.-M. (2018). Adversarial domain adaptation for variational neural language generation in dialogue systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1205–1217.

References III

Tseng, B.-H., Budzianowski, P., Wu, Y.-c., and Gasic, M. (2019). Tree-structured semantic encoder with knowledge sharing for domain adaptation in natural language generation. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 155–164.

Wen, T.-H., Gašić, M., Mrkšić, N., Rojas-Barahona, L. M., Su, P.-H., Vandyke, D., and Young, S. (2016). 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, California. Association for Computational Linguistics.