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*)

```xml
<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>
```
NLG with Iterative Text Editing

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

\[
X_{i-1} = \text{Dublin is the capital of Ireland.} \\
t_i = (\text{Ireland, language, English}) \\
\text{templates} \\
\text{LMScorer} \\
\begin{array}{c|c}
0.8 & 0.9 \\
0.3 & - \\
0.7 & 0.4 \\
\end{array} \\
\text{beam} \\
\text{Dublin is the capital of Ireland, where English is spoken.} = X_i \\
\text{Beam Filter} \\
\text{Sentence Fusion}
\]

- English is spoken in Ireland. 
- One of the languages of Ireland is English.
- English is the official language of Ireland.

Dublin is the capital of Ireland. English is spoken in Ireland.
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 → reinforcement learning?
- Online demo, visualizations
References


