Neural Generation for Czech: Data and Baselines

Ondřej Dušek & Filip Jurčíček
Institute of Formal and Applied Linguistics
Charles University, Prague

INLG, Tokyo, 31 Oct 2019
Task & Motivation

- **Task:** Data-to-text generation from flat MRs
  - as in dialogue systems
  - dialogue act type + attributes/slots + values → sentence in Czech

  inform(name=The Red Lion, food=British)  \[→\]  The Red Lion serves British food.

- **Motivation:** Most data-to-text NLG only targets English
  - non-English systems are mostly handcrafted
  - (surface realization is a different task)

- Not many non-English data-to-text NLG datasets available
- English has little morphology – bias?
- Czech has rich morphology, used in MT a lot, NLP tools ready
Task & Motivation

• Task: Data-to-text generation from flat MRs
  • as in dialogue systems
  • dialogue act type + attributes/slots + values → sentence in Czech

inform(name=Na Růžku, food=Czech) → Na Růžku podávají česká jídla.

• Motivation: Most data-to-text NLG only targets English
  • non-English systems are mostly handcrafted
  • (surface realization is a different task)

• Not many non-English data-to-text NLG datasets available

• English has little morphology – bias?
• Czech has rich morphology, used in MT a lot, NLP tools ready
Delexicalization

• Delexicalization = replacing slot values with placeholders
  • used heavily in NLG systems (not just data-driven)
  • helps fight data sparsity
• Lexicalization = putting concrete values back
  • easy in English – can just do verbatim (for noun phrases)
  • not easy in Czech and other languages with rich morphology
    • need to find the proper surface form to fit the sentence

inform(name=Baráčnická rychta, area=Malá Strana)

Baráčnická rychta  nominative
Baráčnické rychty  genitive
Baráčnické rychtě  dative
Baráčnickou rychtu  accusative
Baráčnické rychtě  locative
Baráčnickou rychtou  instrumental

Malá Strana  nominative
Malé Strany  genitive
Malé Straně  dative
Malou Stranu  accusative
Malé Straně  locative
Malou Stranou  instrumental

\(<name> je na <area>\\n<name> is in <area>\\n<name> najdete v oblasti <area>\\n<name> you-find in the-area of-<area>
Delexicalization

• Delexicalization = replacing slot values with placeholders
  • used heavily in NLG systems (not just data-driven)
  • helps fight data sparsity

• Lexicalization = putting concrete values back
  • easy in English – can just do verbatim (for noun phrases)
  • not easy in Czech and other languages with rich morphology
    • need to find the proper surface form to fit the sentence

```
inform(name=Baráčnická rychta, area=Malá Strana)
```

```
Baráčnická rychta je na Malé Straně
Baráčnická rychta is in Malá Strana

Baráčnickou rychtu najdete v oblasti Malé Strany
Baráčnická rychta you-find in the-area-of Malá Strana
```
Creating a Czech NLG Dataset

- Crowdsourcing was not an option for Czech
  - no Czech speakers on the platforms
- We opted for translating an existing dataset
  - easier than in-house collection
  - translators are easy to hire and require no training
- SFRest (Wen et al., EMNLP 2015)
  - manageable size + shown to work with neural NLG
- We localized the set before translation
  - restaurants, landmarks, addresses in San Francisco → Prague
  - local names sound more natural
  - using various types of names (some inflected, some not)
- We kept track of all possible inflection forms for slot values
Data Statistics

- The result is more complex than SFRest:
  - more distinct lemmas (base forms)
  - >2x more distinct surface word forms
    - not counting restaurant names
  - 3.84 different lexical forms for a slot value on average
- train/dev/test split is not random – we’re ensuring no MR overlap

<table>
<thead>
<tr>
<th></th>
<th>SFRest</th>
<th>CS-Rest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of instances</td>
<td>5,192</td>
<td>5,192</td>
</tr>
<tr>
<td>Unique delexicalized instances</td>
<td>2,648</td>
<td>2,752</td>
</tr>
<tr>
<td>Unique delexicalized MRs</td>
<td>248</td>
<td>248</td>
</tr>
<tr>
<td>Unique lemmas (in delexicalized set)</td>
<td>399</td>
<td>532</td>
</tr>
<tr>
<td>Unique word forms (in delexicalized set)</td>
<td>455</td>
<td>962</td>
</tr>
<tr>
<td>Average lexicalizations per slot value</td>
<td>1</td>
<td>3.84</td>
</tr>
</tbody>
</table>
Model

- Base model: **TGen**
- **Seq2seq with attention**
- **Beam reranking by MR classification**
  - any differences w. r. t. input MR are penalized

- **Base model:**
  - Direct word form generation
  - **Delexicalized input MRs**
**TGen extensions**

- **Lemma-tag generation mode**
  - generate an interleaved sequence of lemmas & morphological tags
  - postprocess using morphological generator (dictionary-based)
  - addressing data sparsity, limiting possible inflection forms for slot values

- **Lexicalized inputs**
  - still generate delexicalized outputs, but input lexicalized MRs
  - some values require different treatment
    - e.g. “in <area>” with different prepositions – *na Smíchově x v Karlíně*
Lexicalization

• New additional generation step
• Baseline: always select most frequent form found in training data
• Non-trivial: RNN LM ranking
  • process sentence up to slot placeholder using LSTM RNN LM
  • get LM probabilities for all possible surface forms for given slot value
  • select the most probable one

```
inform(name=Baráčnická rychta, area=Malá Strana)
```
Lexicalization

• New additional generation step
• Baseline: always select most frequent form found in training data
• Non-trivial: RNN LM ranking
  • process sentence up to slot placeholder using LSTM RNN LM
  • get LM probabilities for all possible surface forms for given slot value
  • select the most probable one

inform(name=Baráčnická rychta, area=Malá Strana)
Evaluation

• BLEU + other E2E metrics
  • single reference → all scores are lower
• Slot error rate (counting placeholders before lexicalization)
• Manually counting errors of different types
  • outputs for each configuration on 100 randomly selected MRs

Results

• Outputs are readable, but not perfect
  • 49% manually evaluated sentences contain some error(s)
  • most problems appear with unusual MRs
Results

- RNN LM for lexicalization helps
  - BLEU improvement statistically significant

- Lexicalized input & lemma-tag help fluency, but hurt accuracy
  - BLEU higher, # fluency errors lower
  - SER + # semantic errors higher

```
<table>
<thead>
<tr>
<th>System configuration</th>
<th>Automatic metrics</th>
<th>Manual evaluation (100 per system)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Generator Mode</td>
<td>Lexicalizer</td>
</tr>
<tr>
<td>Delexicalized</td>
<td></td>
<td>Word forms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RNN LM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lemma-tag</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RNN LM</td>
</tr>
<tr>
<td>Lexicalized</td>
<td></td>
<td>Word forms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RNN LM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lemma-tag</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RNN LM</td>
</tr>
</tbody>
</table>
```
Conclusions

• 1st(?) non-English neural data-to-text NLG dataset + baselines
• Czech harder than English due to slot value inflection
  • using RNN LM for that helps
• Czech may need more data than English

Future work
• pretrain a language model on similar domains
• use MT for synthetic data
Thanks

• Get the code: http://bit.ly/tgen-nlg
• Get the data: http://bit.ly/cs-rest
• Contact me: odusek@ufal.mff.cuni.cz
  @tuetschek

Get this paper: arXiv: 1910.05298
Output examples

**Input DA**
inform(food=Turkish, name="Green Spirit", price_range=expensive)

**Delex. input DAs, any mode**
Green Spirit je dráhý turecká restaurace.

**Most frequent**
is expensive Turkish restaurant

**Delex. input DAs, any mode**
Green Spirit je dráhá turecká restaurace.

**RNN LM**
is expensive Turkish restaurant

**Lex. DAs, word forms**
Green Spirit je dráhý turecká restaurace

**Most frequent**
is expensive Turkish restaurant

**Lex. DAs, word forms**
Green Spirit je dráhá turecká restaurace

**RNN LM**
is expensive Turkish restaurant

**Lex. DAs, lemma-tag**
Green Spirit je dráhý restaurace, která podává turecká jídla.

**Most frequent**
expensive restaurant which serves Turkish meals

**Lex. DAs, lemma-tag**
Green Spirit je dráhá restaurace, která podává turecká jídla.

**RNN LM**
is expensive restaurant which serves Turkish meals

**Input DA**
inform(area=dont.care, count=218, food=dont.care, price_range=dont.care, type=restaurant)

**Delex. input DAs**
Pokud vám nezáleží na druhu jídla, našla jsem 218 restaurace v různých cenových skupinách. *(missing area)*

**Word forms mode**
if you don’t-care about type of-food I-found restaurants in various price ranges

**Delex. input DAs**
Pokud vám nezáleží na druhu jídla, našla jsem 218 restaurací v různých cenových skupinách. *(missing area)*

**Lemma-tag mode**
if you don’t-care about type of-food I-found restaurants in various price ranges

**Lex. input DAs**
V úvahu připadají 218 restaurací, pokud vám nezáleží na druhu jídla. *(missing area)*

**Word forms mode**
into consideration come restaurants if you don’t-care about type of-food if you don’t-care about type of-food *(missing area, price range)*

**Lex. input DAs**
Mám tu 218 restaurací, pokud vám nezáleží na druhu cenových skupinách. *(missing area, food type)*

**Lemma-tag mode**
I-have here restaurants if you don’t-care about type price ranges