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

<name> je na <area>

<name> is in <area>

<name> najdete v oblasti <area>

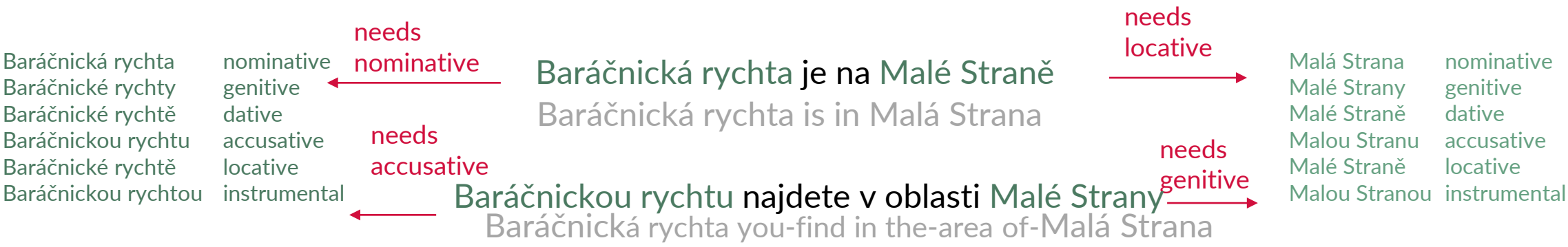
<name> you-find in the-area of-<area>

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

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)



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

Ananta	– feminine, inflected
BarBar	– masculine inanim., inflected
Café Savoy	– neuter, not inflected
Místo	– neuter, inflected
U Konšelů	– prep. phrase, not inflected

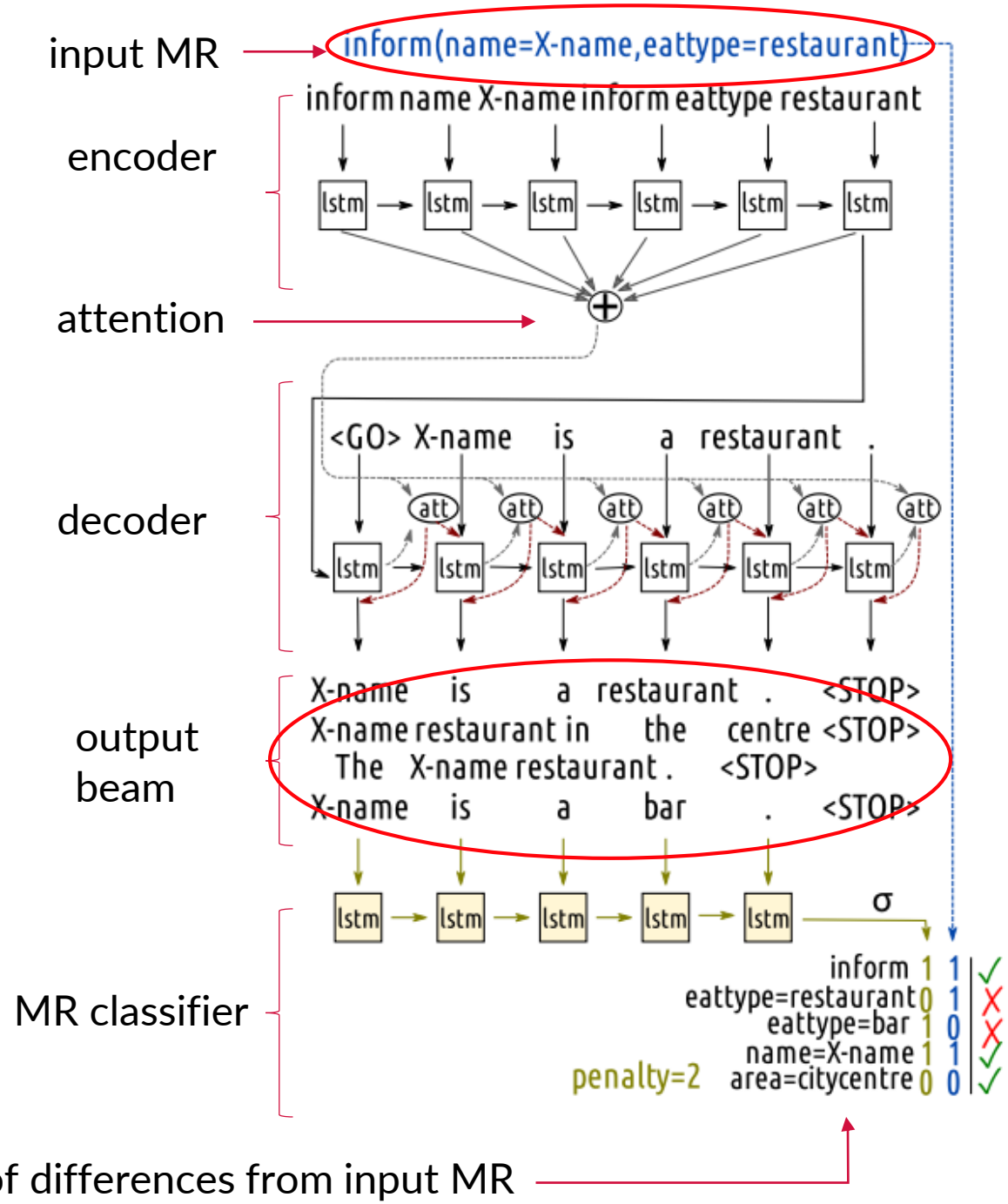
Data Statistics

- 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

	SFRest	CS-Rest
Number of instances	5,192	5,192
Unique delexicalized instances	2,648	2,752
Unique delexicalized MRs	248	248
Unique lemmas (in delexicalized set)	399	**532**
Unique word forms (in delexicalized set)	455	**962**
Average lexicalizations per slot value	1	**3.84**

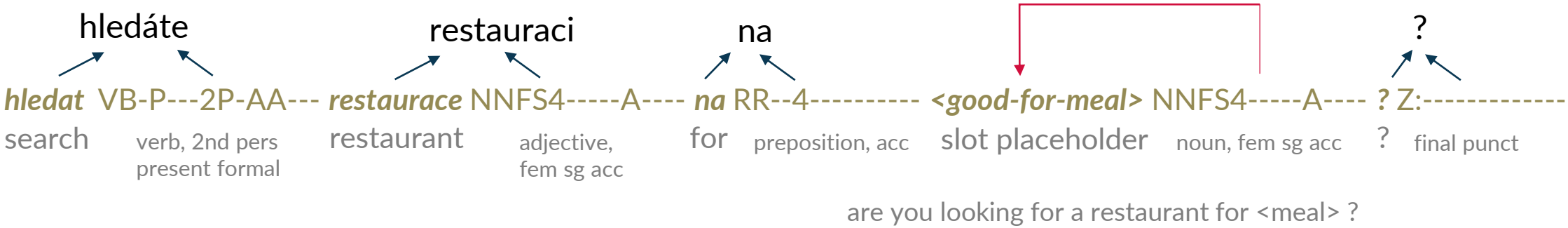
Model

- 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

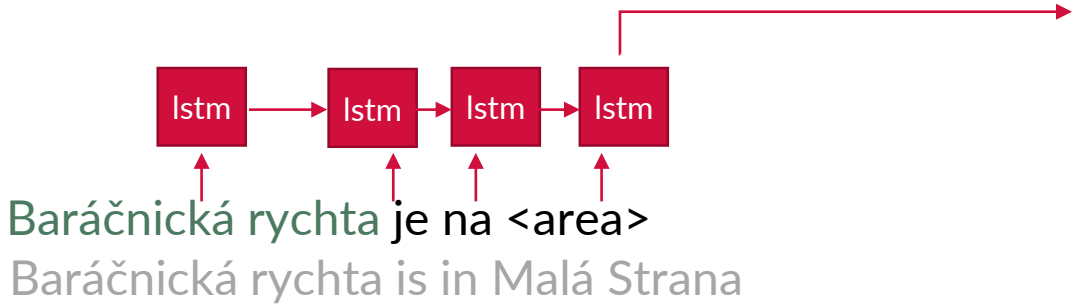


- **Lexicalized inputs**
  - still generate delexicalized outputs, but input lexicalized MRs
  - some values require different treatment
    - e.g. “in <area>” with different prepositions – **na** Smíchově x **v** Karlíně

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)



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

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

System configuration			Automatic metrics			Manual evaluation (100 per system)		
Input DAs	Generator Mode	Lexicalizer	BLEU	NIST	SER	# Semantic Errors	# Repeating Content	# Fluency Errors
Delexicalized	Word forms	Most frequent	20.28	4.519	**0.70**	**8**	**0**	73
		RNN LM	20.74	4.510	**0.70**	**8**	**0**	41
	Lemma-tag	Most frequent	21.21	4.690	1.85	12	2	61
		RNN LM	**21.96**	**4.772**	1.85	12	2	22
Lexicalized	Word forms	Most frequent	19.73	4.562	2.30	14	5	54
		RNN LM	20.48	4.606	2.30	14	5	30
	Lemma-tag	Most frequent	19.44	4.445	3.08	15	4	44
		RNN LM	20.42	4.546	3.08	15	4	**14**

- 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

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
http://bit.ly/odusek
@tuetschek

Get this paper: arXiv: 1910.05298

Output examples

Input DA		*inform(food=Turkish, name="Green Spirit", price_range=expensive)*
Systems	Delex. input DAs, any mode Most frequent	Green Spirit je drahé turecká restaurace. is expensive Turkish restaurant
	Delex. input DAs, any mode RNN LM	Green Spirit je drahá turecká restaurace. is expensive Turkish restaurant
	Lex. DAs, word forms Most frequent	Green Spirit je drahé turecká restaurace is expensive Turkish restaurant
	Lex. DAs, word forms RNN LM	Green Spirit je drahá turecká restaurace is expensive Turkish restaurant
	Lex. DAs, lemma-tag Most frequent	Green Spirit je drahé restaurace, která podává turecká jídla . expensive restaurant which serves Turkish meals
	Lex. DAs, lemma-tag RNN LM	Green Spirit je drahá restaurace, která podává turecká jídla . is expensive restaurant which serves Turkish meals

semantic errors

~~repetition~~

fluency

impossible to lexicalize correctly

lexicalization

Input DA		*inform(area=dont_care, count=218, food=dont_care, price_range=dont_care, type=restaurant)*
Systems (any lexicalizer)	Delex. input DAs Word forms mode	Pokud vám nezáleží na druhu jídla , našla jsem 218 restaurace v různých cenových skupinách. *(missing area)* if you don't-care about type of-food I-found restaurants in various price ranges
	Delex. input DAs Lemma-tag mode	Pokud vám nezáleží na druhu jídla , našla jsem 218 restaurací v různých cenových skupinách. *(missing area)* if you don't-care about type of-food I-found restaurants in various price ranges
	Lex. input DAs Word forms mode	V úvahu připadají 218 restaurací , pokud vám nezáleží na druhu jídla ~~, pokud vám nezáleží na druhu jídla~~ . 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 Lemma-tag mode	Mám tu 218 restaurací , pokud vám nezáleží na druhu cenových skupinách. *(missing area, food type)* I-have here restaurants if you don't-care about type price ranges