OOVs in the Spotlight: How to Inflect them?

Tomáš Sourada, Jana Straková, Rudolf Rosa

{sourada,strakova,rosa}@ufal.mff.cuni.cz

May 2024

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Intro & motivation

Inflection: the task

	input	input	desired output
	Lemma +	**Morphological tag** →	**Inflected form**
en	hug	V;PST	hugged
	spark	V;V.PTCP;PRS	sparking
es	liberar	V;IND;FUT;2;SG	liberarás
	descomponer	V;NEG;IMP;2;PL	no descompongáis
de	aufbauen	V;IND;PRS;2;SG	baust auf
	Ärztin	N;DAT;PL	Ärztinnen
cs	výsledek	N;DAT;SG	výsledku
	analyzovat	V;COND;PRS;2;PL;NEUT	analyzovala byste

Table 1: Example of the inflection task (adapted from Cotterell et al. (2017))

OOVs (out-of-vocabulary words)

- inflecting previously unseen lemma is difficult
- OOV conditions: test lemma not present in the training data
- true OOVs even more difficult: proper nouns, neologisms

Data

There are no OOV data - we need OOV data

- for training
- more importantly for evaluation, to see how well we are performing on OOVs

Czech OOV Inflection Dataset

- large train-dev-test split for standard OOV evaluation
  - auto-extracted from a large morphological dictionary MorfFlex
  - lemma-disjoint (no lemma overlap between the splits)
  - test-MorfFlex (the test set)
- test-neologisms:
  - true OOVs
  - manually annotated set of neologisms
  - evaluation of performance in real-world condition

Czech OOV Inflection Dataset - test-neologisms - example

lemma	tag	form
elektrořidič	S1	elektrořidič
elektrořidič	S2	elektrořidiče
elektrořidič	S3	elektrořidičovi
elektrořidič	S4	elektrořidiče
elektrořidič	S5	elektrořidiči
elektrořidič	S6	elektrořidičovi
elektrořidič	S7	elektrořidičem
elektrořidič	P1	elektrořidiči
elektrořidič	P2	elektrořidičů
elektrořidič	P3	elektrořidičům
elektrořidič	P4	elektrořidiče
elektrořidič	P5	elektrořidiči
elektrořidič	P6	elektrořidičích
elektrořidič	P7	elektrořidiči

Table 2: Example from the test-neologisms dataset: “*elektrořidič*” (driver of an electric car).

Approach

Our 3 approaches

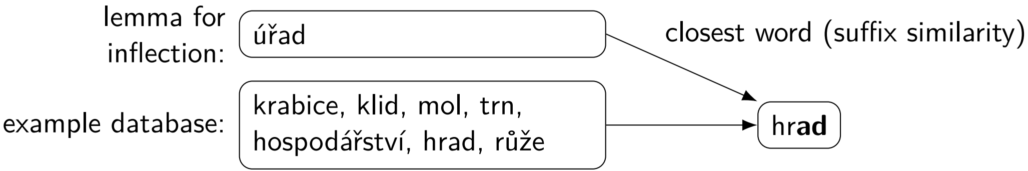
- Retrograde (non-neural approach)
- Seq2seq (encoder-decoder) architecture trained from scratch
    - LSTM
    - Transformer (not a fine-tuned LLM)

Retrograde model

- non-neural approach
- not trained
- inspired in ASIMUT (Králíková and Panevová, 1990)

Retrograde model: how it works

- step 1:



lemma for inflection: úřad

example database: krabice, klid, mol, trn, hospodářství, hrad, růže

closest word (suffix similarity)

hr**ad**

- step 2:

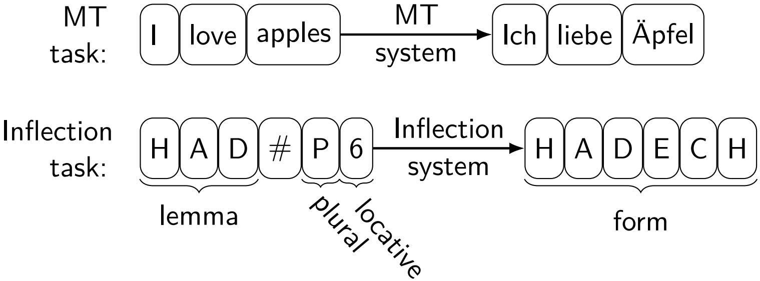
HRAD	
hr-ad	hr-ady
hr-adu	
hr-adu	…
…	
hr-adem	hr-ady

⟶

ÚŘAD	
úř-ad	úř-ady
úř-adu	
úř-adu	…
…	
úř-adem	úř-ady

seq2seq architecture

- LSTM: adapted from Kann and Schütze (2016)
- Transformer: adapted from MT setting
- input-output:



- Surprisingly high batch size needed in the final setup:
  - LSTM: 256
  - Transformer: 4096

Results

SIGMORPHON shared task 2022 results

	Submitted systems					Baselines		Ours	
Lang	CLUZH	Flexica	OSU	TüM	UBC	Neural	NonNeur	LSTM	Transformer
ang	**76.6**	64.4	73.7	71.9	74.1	73.4	68.7	76.3	75.5
ara	81.7	65.5	78.7	78.5	65.5	81.9	50.8	79.2	**82.6**
asm	83.3	75.0	75.0	**91.7**	83.3	83.3	83.3	83.3	83.3
got	92.9	41.4	**94.1**	91.7	91.7	93.5	87.6	92.3	92.3
hun	93.5	62.9	93.1	92.8	91.5	**94.4**	73.1	92.8	**94.4**
kat	96.7	95.7	96.7	96.7	96.7	97.3	96.7	97.3	**97.8**
khk	94.1	47.1	94.1	94.1	88.2	94.1	88.2	**100.0**	94.1
kor	**71.1**	55.4	50.6	56.6	60.2	62.7	59.0	49.4	62.7
krl	87.5	69.8	85.9	57.8	85.4	57.8	20.8	**89.1**	85.9
lud	87.3	92.0	92.9	93.4	88.2	**94.3**	93.4	89.2	92.0
non	85.2	77.0	85.2	80.3	**90.2**	88.5	80.3	83.6	88.5
pol	**96.1**	85.9	94.9	74.0	95.7	74.4	86.3	**96.1**	95.6
poma	76.1	54.5	70.1	69.4	73.3	74.1	47.8	75.2	**76.3**
slk	93.5	90.0	92.2	70.4	**95.7**	71.1	92.4	95.2	**95.7**
tur	93.7	57.9	**95.2**	80.2	92.9	79.4	66.7	**95.2**	92.9
vep	**71.5**	58.8	70.0	57.5	68.8	59.2	60.4	70.7	68.8
average	**86.3**	68.3	83.9	78.6	83.8	80.0	72.2	85.3	86.1

Results on Czech OOV Inflection Dataset

Standard OOV conditions

model	form accuracy	full paradigm accuracy
rule-based sklonuj.cz	88.88	74.43
SIGMORPHON nonneural	94.78	88.15
SIGMORPHON transformer	95.47	87.29
Retrograde	94.85	88.64
LSTM	96.16	89.80
TRM	**96.18**	**90.44**
Upper bound (oracle)	99.3	97.3

Results on Czech OOV Inflection Dataset

True OOVs: neologisms

model	form accuracy	full paradigm accuracy
RULE-BASED SKLONUJ.CZ	86.22	55
SIGMORPHON NONNEURAL	**89.49**	**71**
SIGMORPHON TRANSFORMER	87.53	63
RETROGRADE	89.34	**71**
LSTM	86.95	58
TRM	87.24	61

Summary

1. Transformer works the best in standard OOV conditions
2. But on on true OOVs (neologisms), it is beaten by the retrograde model.
3. Release: Czech OOV Inflection Dataset, ready-to-use inflection library
4. Discussion challenge:
    - small test set: would it scale on a large one?
    - train data to inflect OOVs - how?

See you @ LREC-COLING 2024

References I

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In Mans Hulden, editor, *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-2001. URL `https://aclanthology.org/K17-2001`.

Katharina Kann and Hinrich Schütze. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 555–560, Berlin, Germany, August 2016. Association for Computational Linguistics. URL `http://anthology.aclweb.org/P16-2090`.

Květoslava Králíková and Jarmila Panevová. ASIMUT - a method for automatic information retrieval from full texts. *Explizite Beschreibung der Sprache und automatische Textbearbeitung*, XVII, 1990.

Acknowledgements

Computational resources for this work were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

The work described herein uses resources hosted by the LINDAT/CLARIAH-CZ Research Infrastructure (projects LM2018101 and LM2023062, supported by the Ministry of Education, Youth and Sports of the Czech Republic).

This work has also been supported by the Grant Agency of the Czech Republic under the EXPRO program as project "LUSyD" (project No. GX20-16819X).

We thank the Bernard Bolzano Endowment Fund for the contribution for covering the travel expenses to present this work.

We also thank the anonymous reviewers for their valuable comments.