Transformers: shining in **OOV inflection**, but **beaten on** true **neologisms** by a non-neural model

OOVs in the Spotlight: How to Inflect them?

Tomáš Sourada, Jana Straková, Rudolf Rosa

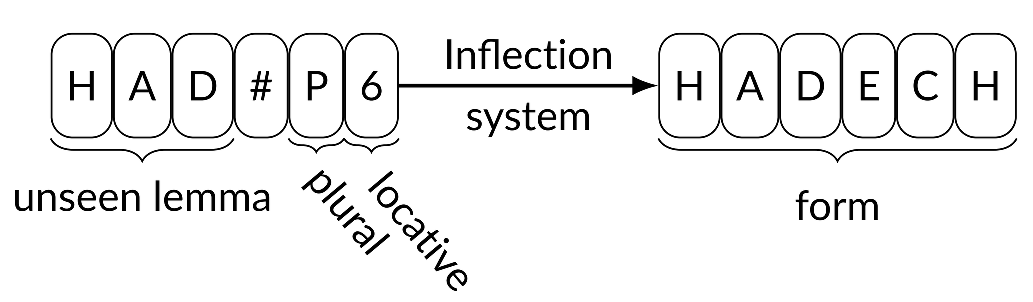
`{sourada,strakova,rosa}@ufal.mff.cuni.cz`

Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic

Introduction

- task:

  lemma + number + case → inflected form



- morphological inflection in out-of-vocabulary (OOV) conditions
- SOTA systems may struggle when asked to inflect unseen lemma
- focus on Czech nouns, extend to 16 languages and all POS

Data

- train-dev-test split of morphological dictionary (lemma-disjoint)
- additional test set of neologisms (real-world OOVs)
- released as Czech OOV Inflection Dataset

Methods

- non-neural retrograde model (based on suffix similarity of lemmas)
- LSTM- and Transformer-based encoder-decoder, trained from scratch, extensively tuned

Results

- SIGMORPHON'22 OOV conditions:
  - our seq2seq SOTA in 9 of 16 langs
- Czech simulated OOV conditions:
  - Transformer performs the best
- Czech true OOVs (neologisms):
  - overall performance drop
  - Transformer defeated by non-neural models

Discussion

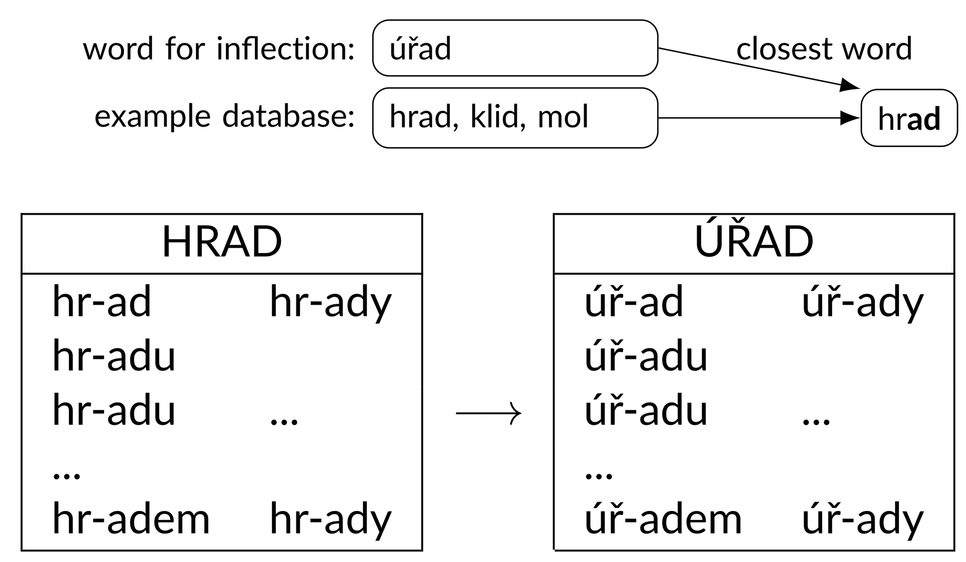
- small test set of neologisms: would the results scale on a large one?
- is morphological dictionary a good train set for inflecting true OOVs?
- if not, how to select the train data?

Datasets

lemma	tag	form
elektrořidič	S1	elektrořidič
elektrořidič	S2	elektrořidiče
elektrořidič	S3	elektrořidičovi
	…	
elektrořidič	P6	elektrořidičích
elektrořidič	P7	elektrořidiči

Set	lemmas	forms	Source
train	360k	5.04M	MorfFlex
dev	44k	616k	MorfFlex
test-MorfFlex	44k	616k	MorfFlex
test-neologisms	101	1.4k	Čeština 2.0

Non-neural retrograde model



Results

test set	MorfFlex		neologisms	
model	FA	FPA	FA	FPA
`copy`	22.6	1.5	13.1	0
`sklonuj.cz`	88.9	74.4	86.2	55
`SIG nonneur`	94.8	88.2	**89.5**	**71**
`SIG trm`	95.5	87.3	87.5	63
`SIG trm-tune`	96.2	90.2	86.5	55
`Retrograde`	94.9	88.6	89.3	**71**
`LSTM`	96.2	89.8	87.0	58
`Transformer`	**96.2**	**90.4**	87.2	61
`majority-vote`	96.4	90.7	90.4	64

SIGMORPHON 2022 comparison

	Submitted systems					Baselines		Ours	
Lang	CLUZH	Flexica	OSU	TüM	UBC	Neural	NonNeur	LSTM	TRM
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

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. This template is based on Mike Morrison's idea of Better Poster. It was modified to latex by Rafael Bailo and modified for UFAL purposes by Tom Kocmi.