

DMoG: A Data-Based Morphological Guesser aka já buřtgulám, ty buřtguláš, vřichni chceme buřtgulat

Vojtěch Kovář, Pavel Rychlý

Natural Language Processing Centre
Faculty of Informatics, Masaryk University

Lexical Computing

{xkovar3,pary}@fi.muni.cz

RASLAN 2021

Morphological guesser use cases

- Lemmatizing out-of-vocabulary (OOV) words
 - buřtguľáš, online, komorbidity, flash, groupe, knedlo, nVidia
- Bootstrapping lemmatization of a new language
 - manually annotate part of the word list
 - learn patterns
 - automatically annotate next part
 - manually fix the annotations
 - learn better patterns
 - repeat until happy

Existing approaches

- CSTlemma, Šmerk 2008 (desamb)
 - based on matching affixes
 - the longer match, the better
 - guessing one OOV word at a time

Existing approaches (II)

■ Funny effects (Czech)

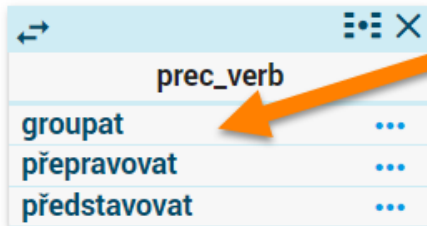
- buřtguľáš → buřtguľat
- knedlo → knednout
- flash → flasha
- nVidia → nVidium
- online → onlinout
- komorbiditou → komorbiditý (adj)
- knedlo → knednout

Existing approaches (III)

WORD SKETCH

EUR-Lex Czech 2/2016

pes as 30,941× ...



↔		☰ ☱ ✕
prec_verb		
groupat	...	
přepřavovat	...	
představovat	...	

Existing approaches (III)

CONCORDANCE

EUR-Lex Czech 2/2016



CQL pes + groupat • 169

0.34 per million tokens • 0.000034%

 Details

Left context

KWIC

Right context

1	<input type="checkbox"/>		engagées à notifier un plan de restructuration du groupe	PSA	Peugeot Citroën S.A. (C
2	<input type="checkbox"/>		it Citroën S.A. (ci-après "PSA" ou le "groupe" ou le " groupe	PSA	") et un plan de viabilité
3	<input type="checkbox"/>		fié le 12 mars 2013 un plan de restructuration du groupe	PSA	ainsi qu'un plan de viab
4	<input type="checkbox"/>		DESCRIPTION DES FAITS </s><s> 2.1. </s><s> Le groupe	PSA	</s><s> (5) </s><s> Le
5	<input type="checkbox"/>		1. </s><s> Le groupe PSA </s><s> (5) </s><s> Le groupe	PSA	est une société cotée s
6	<input type="checkbox"/>		</s><s> Présent commercialement dans 160 pays, le groupe	PSA	exploite 11 usines dite

- groupe → groupat

- *A kde je babička? — Ále, zase groupe psa.*

Our approach

- Organize affixes into groups/patterns
- For each candidate word-form → lemma
 - generate all forms predicted by the pattern
 - check how many of them are present in the corpus (word list)
 - pattern with most predictions present in the corpus is the best
- I.e. process the whole corpus word list as the input
 - instead of isolated word forms

Example: Buřtguľáš

- Buřtguľáš → buřtguľat (verb) ? (like *děľáš*)
 - predicts buřtguľám, buřtguľáš, buřtguľá, buřtguľáme, ...
 - only **buřtguľáš** is present in the corpus
- Buřtguľáš → buřtguľáš (noun) ? (like *mariáš*)
 - predicts buřtguľáše, buřtguľášem, buřtguľáši, buřtguľášů, ...
 - some of them will be present in the corpus
 - better candidate

Implementation

■ Prototype implementation in Python

- only suffixes
- no connection to PoS categories
- simple: two scripts (train, evaluate), <120 lines of code
- can be extended easily

■ Pattern

- set of suffix pairs
- $\{(-ám, -at), (-áš, -at), (-á, -at), (-ál, -at), (-l, -t), (-jí, -t), \dots\}$

Evaluation

- Extremely preliminary :)
- 40 most frequent OOV words from csTenTen17 web corpus
- Patterns trained on DESAM manually curated corpus
- Results
 - correct: 36, incorrect: 4, accuracy: 90 %
- Compared with Šmerk (2008)
 - correct: 26, incorrect: 14, accuracy: 65 %

Conclusions

- New method of lemmatization for OOV words
- Promising initial results, more work needed