MUNI
FI

A new Czech pipeline

Vlasta Ohlídalová
vlasta.ohlidalova@sketchengine.eu

Faculty of Informatics, Masaryk University

December 7, 2024

Pipeline in this presentation

Mám chuť na pierogy s cukrem.

```
Mám      k5eAaImIp1nSrD  mít-v    mít
chuť     k1gFnSc4        chuť-n   chuť
na       k7c4     na-p   na
pierogy  k1c4gMnP        pierogy-n        pierogy
s        k7c7     s-p    s
cukrem   k1gInSc7        cukr-n   cukr
<g/>
.        kIxD     .-x    .
```

The original pipeline...

normalization	tokenization	morphological analyzer	guesser	POS tagging
↓	↓	↓	↓	↓
uninorm	unitok	Majka	guesser	desamb

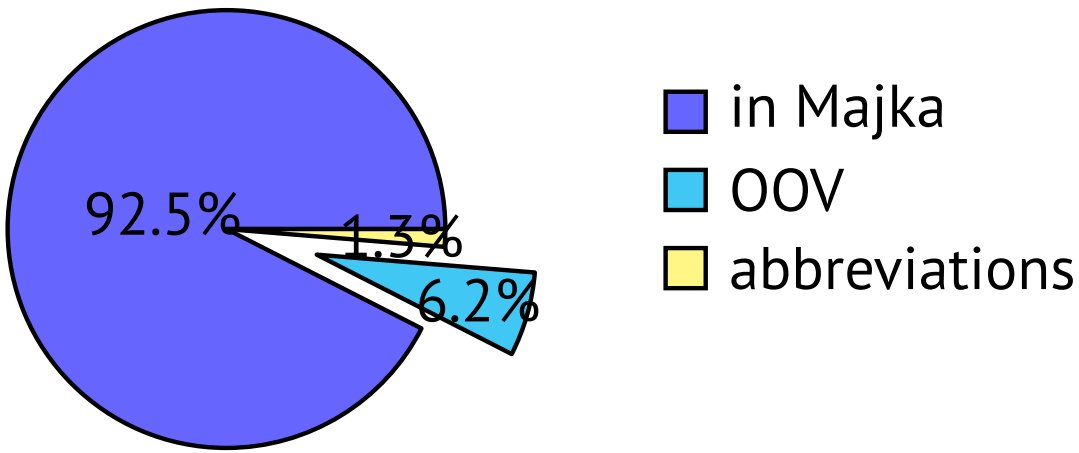
...and the current one.

normalization	tokenization	morphological analyzer	guesser	POS tagging
↓	↓	↓	↓	↓
uninorm	unitok	Majka	guesser	RFTagger

Majka

Czech dictionary headwords coverage in Majka

- Currently 61,676 lemmas approved in the project



Majka

The missing words

word	order	frequency
online	478	834843
info	2983	164701
Wi-Fi	4395	114165
blog	4625	108700
iPhone	6133	82655
nej	7151	69862
CO2	7248	68886
fitness	7429	67087
wellness	7864	62956
play-off	10320	47187
D1	10967	44251
e-shop	11456	42226
dovolatel	14758	31883
live	15083	31123
make-up	15268	30680

Majka

The missing words

word	order	frequency
online	478	834843
info	2983	164701
Wi-Fi	4395	114165
blog	4625	108700
iPhone	6133	82655
nej	7151	69862
CO2	7248	68886
fitness	7429	67087
wellness	7864	62956
play-off	10320	47187
D1	10967	44251
e-shop	11456	42226
dovolatel	14758	31883
live	15083	31123
make-up	15268	30680

DESAM

POS tagging training data

- incomplete annotation,
- multi-word expressions,
- discrepancy between Majka and DESAM

DESAM

POS tagging training data

- **incomplete annotation**,
- multi-word expressions,
- discrepancy between Majka and DESAM

DESAM

POS tagging training data

- incomplete annotation,
- **multi-word expressions**,
- discrepancy between Majka and DESAM

Table: Multi-word expressions in DESAM

Na rozdíl od
Mimo jiné
Z hlediska
V důsledku
V souvislosti s
V porovnání s
Karlovy Vary
Las Vegas
Australian Open
Tour de France
i když

DESAM

POS tagging training data

- incomplete annotation,
- multi-word expressions,
- **discrepancy between Majka and DESAM**.
  1. však [conjunction, particle] → [conjunction]
  2. už [particle]→ [adverb, verb]
  3. manual annotation
  4. out-of-vocabulary words → noun

DESAM

Changes overview

Table: The most frequently changed tokens in DESAM.

frequency	word	original tag	new tag
508	ještě	k9	k6eAd1
575	a	k9	k8xC
579	tedy	k9	k8xC
625	totiž	k9	k8xC
654	proto	k8xC	k6eAd1xC
686	tak	k9	k6eAd1tMtQxCxD
705	jako	k9	k8xS
778	jak	k8xS	k6eAd1yR
822	až	k9	k8xS
1,225	také	k9	k6eAd1
2,203	však	k9	k8xC
3,662	i	k9	k8xC

RFTagger evaluation

- Evaluation on kgncp (POS, gender, number, case, person) only,
- tagset modifications for the final version.

	original %	Majka %
error	8.616	7.462
error in kgncp	**8.279**	**7.118**
k	2.121	0.984
c	3.886	3.878
g	1.991	1.969
n	1.179	1.125
p	0.002	0.001

RFTagger tagset modifications

The main idea: tagset modifications should never lead to losing information.

- Training on specific attributes only,
- change order of the attributes,
- special attribute for "být",
- proper nouns tagged,
- passive participle (verb –> adj).

Pipeline in practice

```
[Saturday  7 December 2024 14:43][vlasta@toad8 /mnt/toad8_disk1/vlasta/czech_pipeline][0]$
>vim guesser.py
[Saturday  7 December 2024 14:45][vlasta@toad8 /mnt/toad8_disk1/vlasta/czech_pipeline][0]$
>cat /tmp/mamchut | /mnt/toad8_disk1/vlasta/majka/majka -t -p -f /mnt/toad8_disk1/vlasta/lemma_fsa/czech/w-l
mám     máma:k1.c2.gF.nP.-/k1gFnPc2     mít:k5.-.nS.-.p1.mI/k5eAaImIp1nSrD
chuť    chuť:k1.c1.gF.nS.-/k1gFnSc1     chuť:k1.c4.gF.nS.-/k1gFnSc4
na      na:k7.c4/k7c4   na:k7.c6/k7c6
pierogy
s       s:k0/k0 s:k7.c7/k7c7    s:kA/kA
cukrem  cukr:k1.c7.gI.nS.-/k1gInSc7
.
[Saturday  7 December 2024 14:53][vlasta@toad8 /mnt/toad8_disk1/vlasta/czech_pipeline][0]$
>cat /tmp/majka | python guesser.py -d affix.dic -c cstenten_all_mj2
mám     máma:k1.c2.gF.nP.-/k1gFnPc2     mít:k5.-.nS.-.p1.mI/k5eAaImIp1nSrD
chuť    chuť:k1.c1.gF.nS.-/k1gFnSc1     chuť:k1.c4.gF.nS.-/k1gFnSc4
na      na:k7.c4/k7c4   na:k7.c6/k7c6
pierogy pierogi:k1.c4.gM.nP.xF/k1gMnPc4xF
s       s:k0/k0 s:k7.c7/k7c7    s:kA/kA
cukrem  cukr:k1.c7.gI.nS.-/k1gInSc7
.
[Saturday  7 December 2024 14:53][vlasta@toad8 /mnt/toad8_disk1/vlasta/czech_pipeline][0]$
>cat /tmp/guesser | ../rftagger/src/rft-annotate -x -q -a czech.par
No plausible tags for pierogy after filtering by input analysis => discarding input analysis
mám     mít     k5eAaImIp1nSrD
chuť    chuť    k1gFnSc4
na      na      k7c4
pierogy k1.c4.gM.nP.-
s       s       k7c7
cukrem  cukr    k1gInSc7
.       kI.xD
```

RFTagger evaluation

- Evaluation on kgncp only,
- tagset modifications for the final version.

	original %	Majka %	changed %
error	8.616	7.462	6.848
error in kgncp	**8.279**	**7.118**	**6.293**
k	2.121	0.984	0.786
c	3.886	3.878	3.31
g	1.991	1.969	1.926
n	1.179	1.125	1.058
p	0.002	0.001	0.001

Known issues & future work

- Lemma disambiguation,
- guesser.

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS