Neural Tagger for Czech Language: Capturing Linguistic Phenomena in Web Corpora

Zuzana Nevěřilová, Marie Stará

Natural Language Processing Centre
Masaryk University

December 6, 2019

Tagging Czech

Known Issues in majka + guesser + desamb

Neural Tagger for Czech

Evaluation

Example Outputs

Side Effects

Tagging Czech . . .

. . . and other languages with rich inflection

- challenging because of the number of possible tags
  In Universal dependencies:
  - ajka tagset: 2,176 tags
  - Penn Tree Bank tagset: 48 tags

Tagging Czech . . .

. . . and other languages with rich inflection

- challenging because of the number of possible tags
  In Universal dependencies:
  - `ajka` tagset: 2,176 tags
  - Penn Tree Bank tagset: 48 tags
- challenging because of foreign words

Tagging Czech . . .

. . . and other languages with rich inflection

- challenging because of the number of possible tags
  In Universal dependencies:
  - `ajka` tagset: 2,176 tags
  - Penn Tree Bank tagset: 48 tags
- challenging because of foreign words
- challenging because of multi-word expressions that do not follow the syntactic rules (and often contain foreign words)

Current Taggers for Czech

- Morče → MorphoDiTa [Straková et al., 2014] (95.75% accuracy in POS tagging)
- MUMULS [Variš and Klyueva, 2018], MWE-aware tagger for 10 languages (incl. Czech)
- `majka` + guesser + `desamb`, almost no documentation but known issues

Known Issues in majka + guesser + desamb

Strategickými	partnery	jsou	Trust	for	Civil
ADJ	NOUN	VERB	NOUN	NOUN	NOUN
k2eAgMnPc7d1	k1gMnPc7	k5eAaImIp3nP	k1gInSc4	k1gNnPc2	k1gMnSc1
strategický	partner	být	trust	forum	civil
Society	in	Central	and	Eastern	Europe
NOUN	?	NOUN	?	NOUN	NOUN
k1gFnSc2	k?	k1gMnSc1	k?	k1gMnSc1	k1gMnSc5
societa	in	Central	and	Eastern	Europ
,	Google	a	Samsung	Electronics	.
PUNCT	NOUN	CONJ	ABBR	NOUN	PUNCT
kIx,	k1gMnSc5	k8xC	kA	k1gInSc1	kIx.
,	Googl	a	Samsung	Electronics	.

features: foreign words, named entities, MWEs

errors: incorrect case, forced gender assignment

Known Issues in majka + guesser + desamb

Jejich	kůže	je	žlutohnědá	nebo	šedá	a
PRON	NOUN	VERB	ADJ	CONJ	VERB	CONJ
k3xOp3gInPc4	k1gFnSc1	k5eAaImIp3nS	k2eAgFnSc1d1	k8xC	k5eAaImIp3nS	k8xC
jejich	kůže	být	žlutohnědý	nebo	šedat	a
jejich	vlasy	jsou	tmavě	hnědé	nebo	černé
PRON	NOUN	VERB	ADV	ADJ	CONJ	ADJ
k3xOp3gInPc1	k1gInPc1	k5eAaImIp3nP	k6eAd1	k2eAgFnPc1d1	k8xC	k2eAgFnPc1d1
jejich	vlas	být	tmavě	hnědý	nebo	černý

features: isolated adjectives

errors: wrong POS, nominative vs. accusative

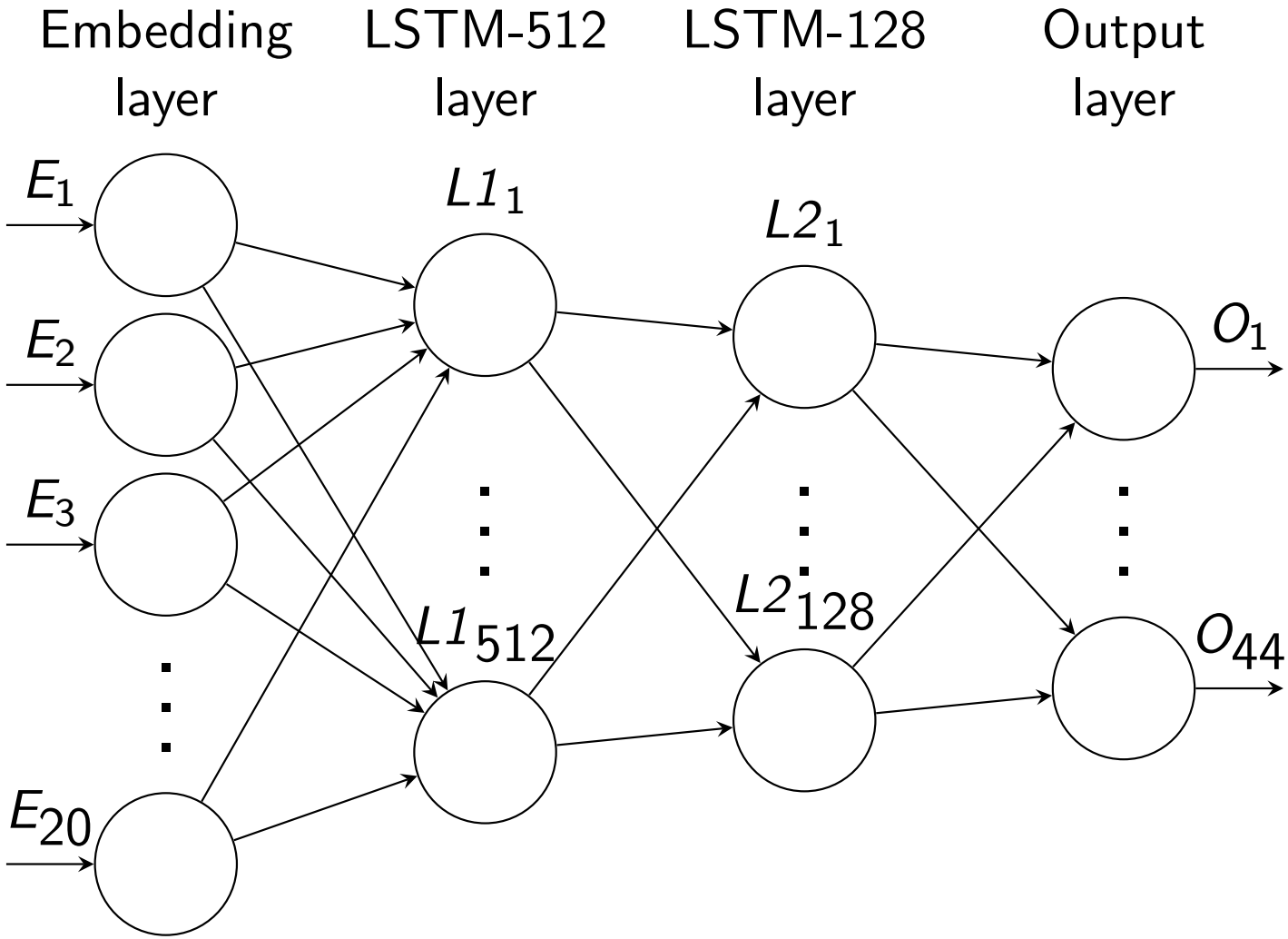
Neural Tagger for Czech

- trained on nice examples from the current CsTenTen corpus [Suchomel, 2018]
- for experimental reasons the tagset was slightly reduced
- the sentence length was limited to 20 tokens
- uses pretrained word embeddings (fasttext [Bojanowski et al., 2016])
- problem reformulation: instead of binary classification in > 1,000 classes (for each tag), we use multinomial classification in 44 classes

Example

`k1gInSc1`	`k1, gI, nS, c1`
tag	attributes

Neural Network Architecture



Training Details

- 9 epochs
- 180,000 samples (171,000/19,000 train/test data)
- validation set accuracy: 99.38%
- validation loss: 1.70%
- model size: 18MB (+ 7GB fasttext model)

The optimistic accuracy is caused by the zero padding.

Evaluation

- quantitative: how many tags were correct?
- qualitative: what are the most common errors and how serious are they?

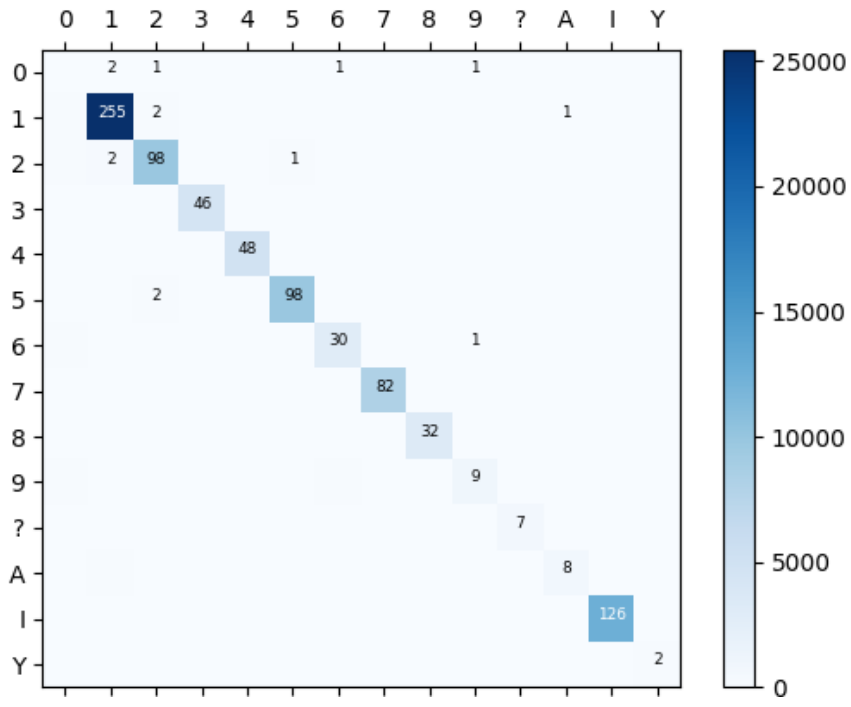
Quantitative Evaluation

evaluation on 6,950 sentences

- 75.25% exact match
- 87.62% submatch (the predicted tag is contained in the golden tag)
- 91.62% match on attributes
- 96.5% match on POS attribute
- 1.2% no tag predicted

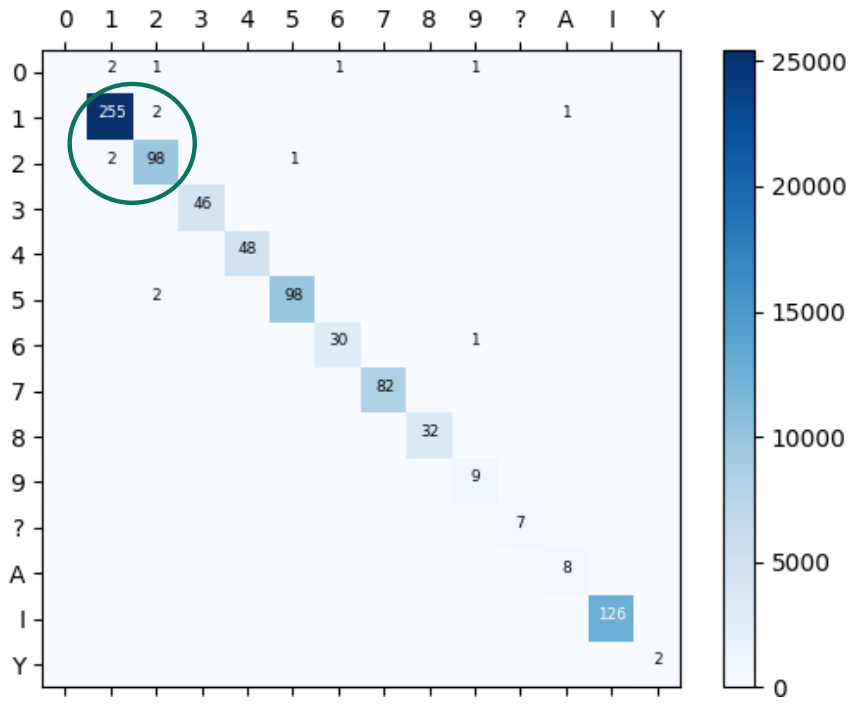
Qualitative Evaluation

the most serious problem is wrong POS tag:



Qualitative Evaluation

the most serious problem is wrong POS tag:

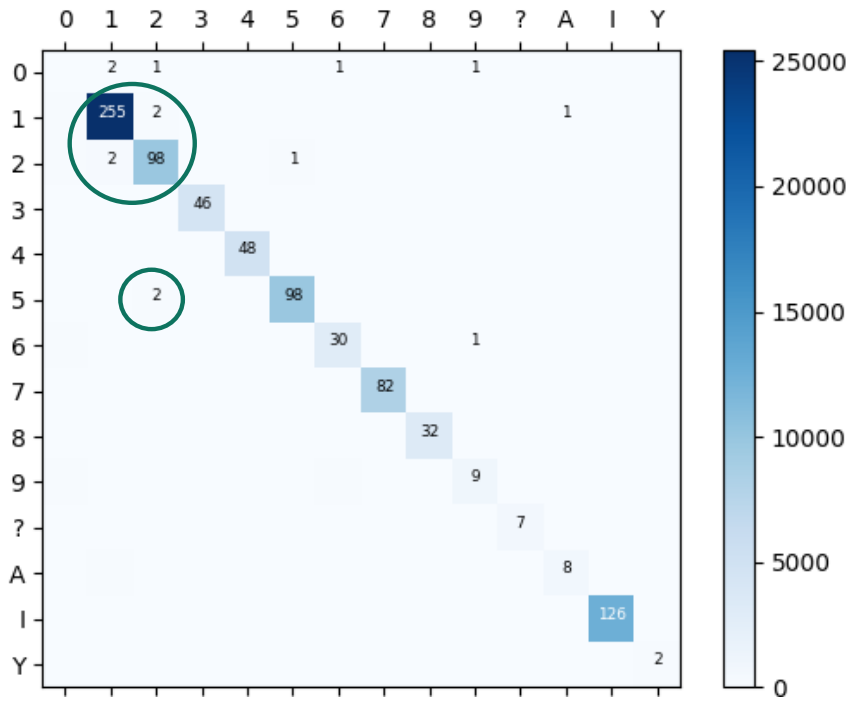


nouns and adjectives

svíčková, internet banking

Qualitative Evaluation

the most serious problem is wrong POS tag:

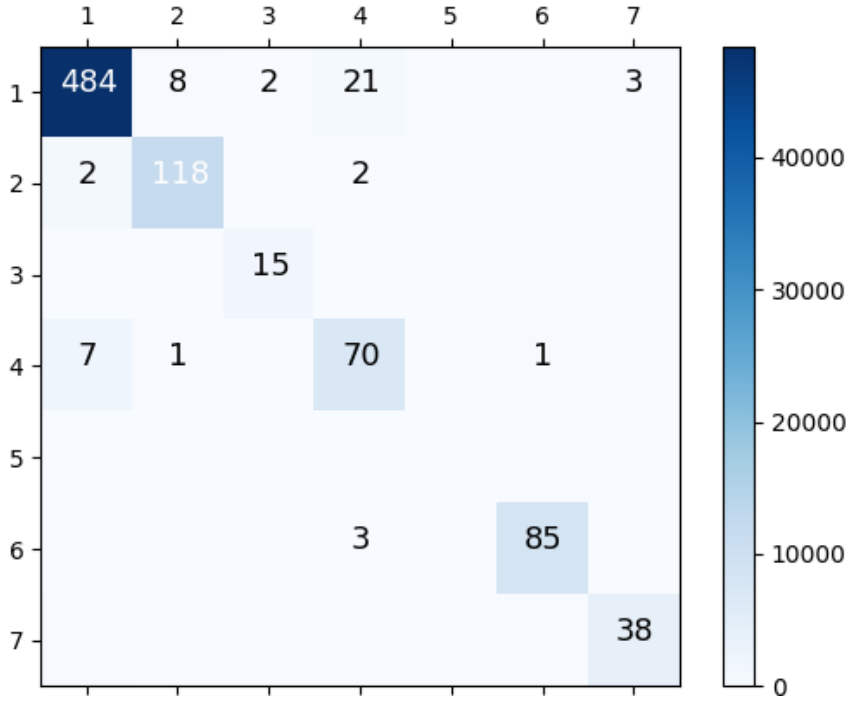


nouns and adjectives
svíčková, internet banking

verbs and adjectives
mlád, informován

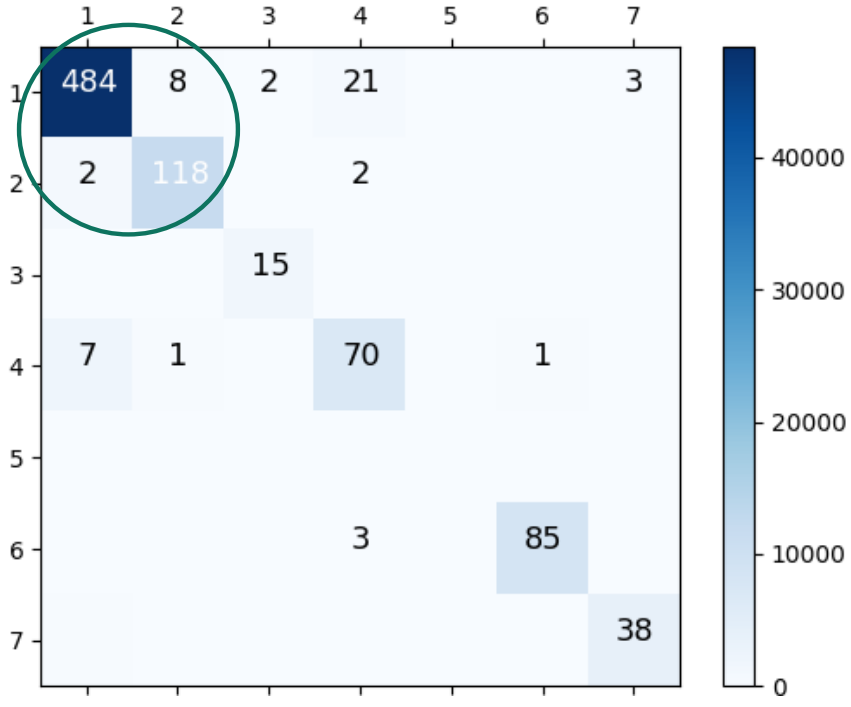
Qualitative Evaluation

the second most serious problem is wrong case tag:



Qualitative Evaluation

the second most serious problem is wrong case tag:



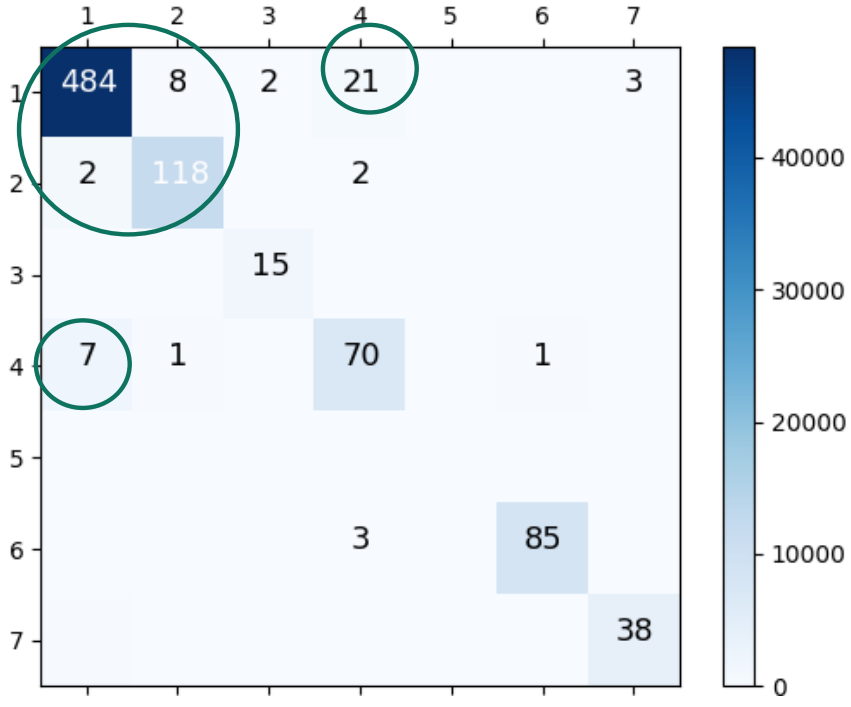
nominative and genitive

výstavba základny údržby

Tom Petty

Qualitative Evaluation

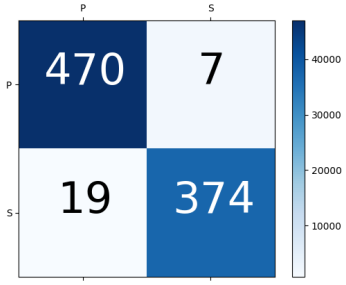
the second most serious problem is wrong case tag:



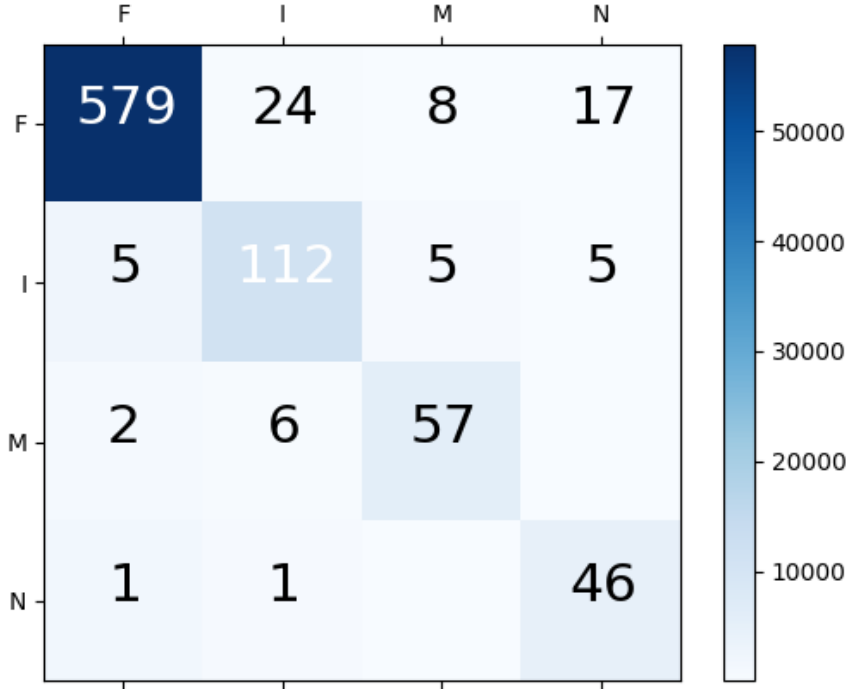
nominative and genitive
výstavba základny údržby
Tom Petty

nominative and accusative

Qualitative Evaluation



number



gender

Example Outputs

Strategickými	partnery	jsou	Trust	for	Civil
ADJ	NOUN	VERB	NOUN	NOUN	NOUN
k2gMnPc7d1	k1gMnPc7	k5mIp3nP	k1gInSc4	k1gNnPc2	k1gMnSc1
k2gInPc7d1	k1nPc7	k5mIp3nP	k1nSc1	nS	k1nS
strategický	partner	být	trust	forum	civil
Society	in	Central	and	Eastern	Europe
NOUN	?	NOUN	?	NOUN	NOUN
k1gFnSc2	k?	k1gMnSc1	k?	k1gMnSc1	k1gMnSc5
k1nS	k?	k1nSc1	k1nSc1	k1nSc1	k1nS
societa	in	Central	and	Eastern	Europ
,	Google	a	Samsung	Electronics	.
PUNCT	NOUN	CONJ	ABBR	NOUN	PUNCT
kIxX	k1gMnSc5	k8	kA	k1gInSc1	kIxX
kIxX	k1nSc1	k8	k1nSc1	k1nSc1	kIxX
,	Googl	a	Samsung	Electronics	.

Example Outputs

Jejich	kůže	je	žlutohnědá	nebo	šedá	a
PRON	NOUN	VERB	ADJ	CONJ	VERB	CONJ
k3p3gInPc4	k1gFnSc1	k5mIp3nS	k2gFnSc1d1	k8	k5mIp3nS	k8
k3p3gFnSc1	k1gFnSc1	k5mIp3nS	k2gFnSc1d1	k8	k2gFnSc1d1	k8
jejich	kůže	být	žlutohnědý	nebo	šedat	a
jejich	vlasy	jsou	tmavě	hnědé	nebo	černé
PRON	NOUN	VERB	ADV	ADJ	CONJ	ADJ
k3p3gInPc1	k1gInPc1	k5mIp3nP	k6d1	k2gFnPc1d1	k8	k2eAgFnPc1d1
k3p3gInPc1	k1gInP	k5mIp3nP	k6d1	k2nPc1d1	k8	k2gFnPc1d1
jejich	vlas	být	tmavě	hnědý	nebo	černý

Side Effects of Using fasttext

- no OOV problem since fasttext contains subword information
- implicit information about frequent collocations from fasttext → MWE annotation

From the above examples, the following MWEs were detected: Eastern Europe, Samsung Electronics, tmavě hnědé

	tmavě	hnědé	tmavěhnědé
tmavě	1.	0.56104434	0.64970
hnědé	0.561044	1.	0.706171
tmavěhnědé	0.64970	0.706171	1.

cosine similarity matrix for w_1, w_2, and $concat(w_1, w_2)$

Future Work

Future Work

- better training data, cleaning
- revert the reduced tagset
- deploy as a service
- generate lemmata (and replace guesser)

Future Work

- better training data, cleaning
- revert the reduced tagset
- deploy as a service
- generate lemmata (and replace guesser)

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016).
Enriching word vectors with subword information.
arXiv preprint arXiv:1607.04606.

Straková, J., Straka, M., and Hajič, J. (2014).
Open-source tools for morphology, lemmatization, POS tagging and named entity recognition.
In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Suchomel, V. (2018).
csTenTen17, a recent czech web corpus.
Twelveth Workshop on Recent Advances in Slavonic Natural Language Processing, pages 111–123.


Variš, D. and Klyueva, N. (2018).
Improving a neural-based tagger for multiword expressions identification.
In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).