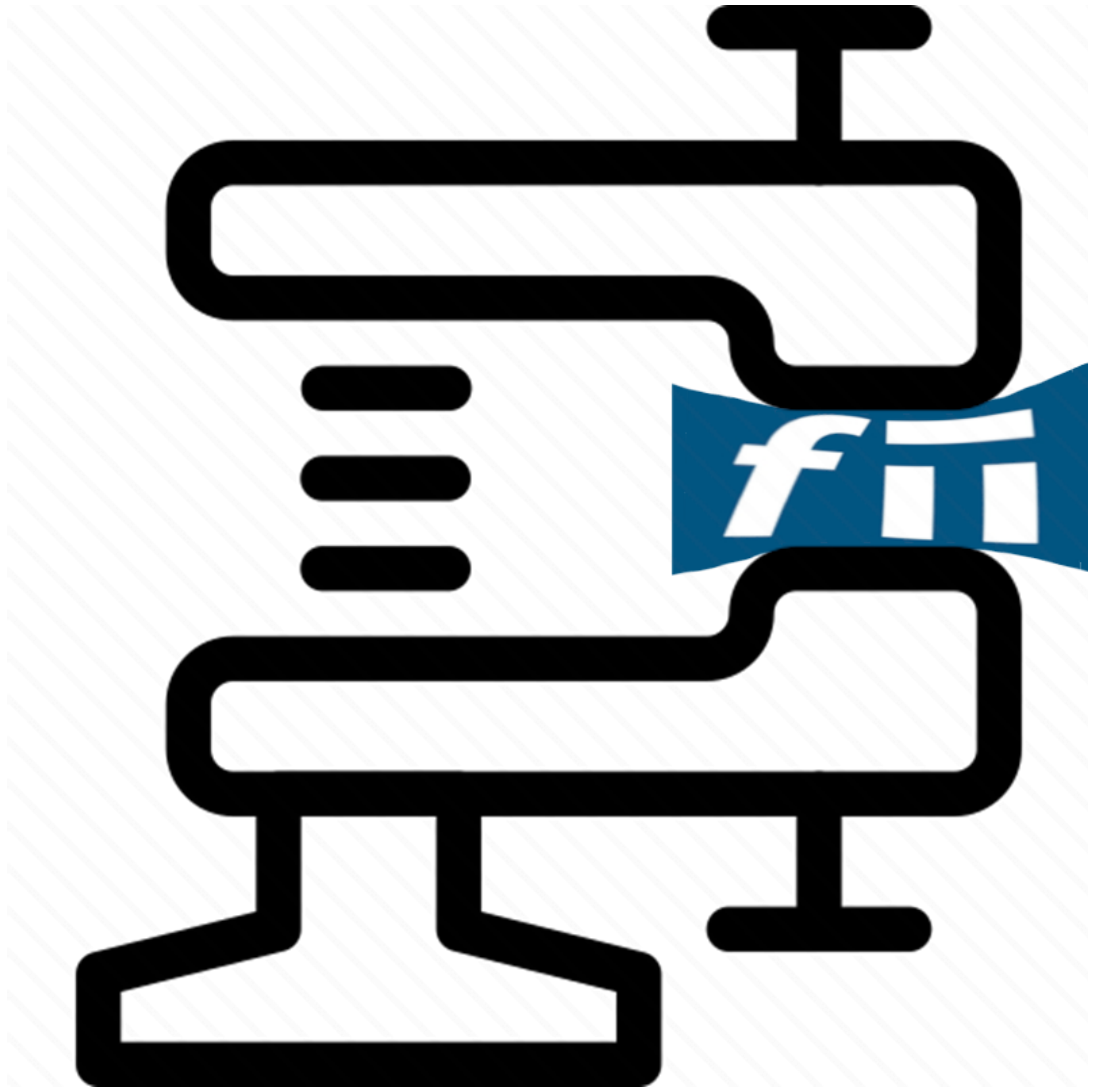


COMPRESSED FASTTEXT MODELS

Zuzana Nevěřilová | RASLAN 2022
December 9, 2022 | Karlova Studánka

TRUE STORY

I’m not happy with our tagger.
Make some better.
...
Here it is.
Ten-gigabyte server? We cannot deploy this.
...
What about 1.7GB?

AGENDA

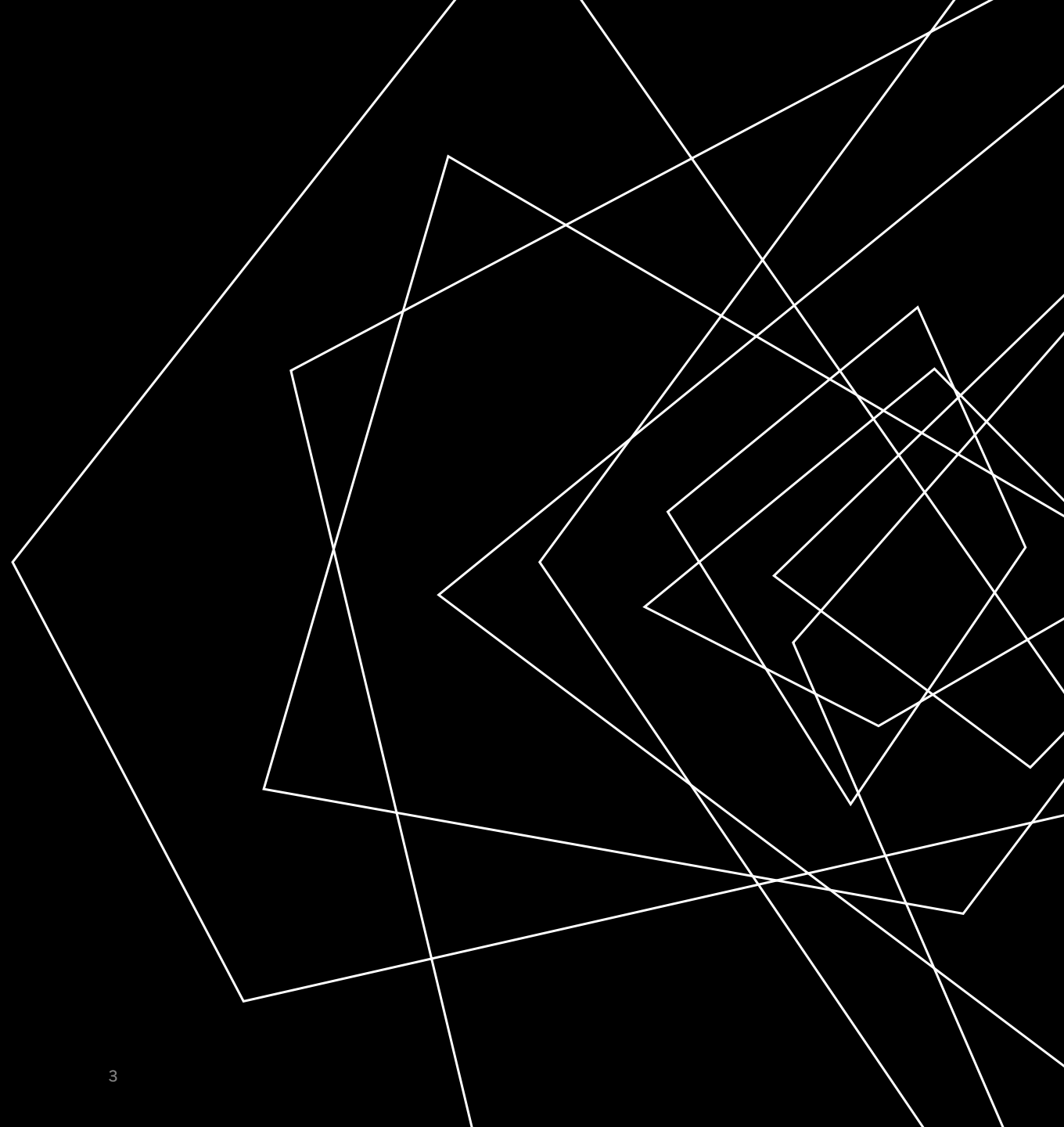
Word Embeddings

Compression Methods

Tagger for Czech

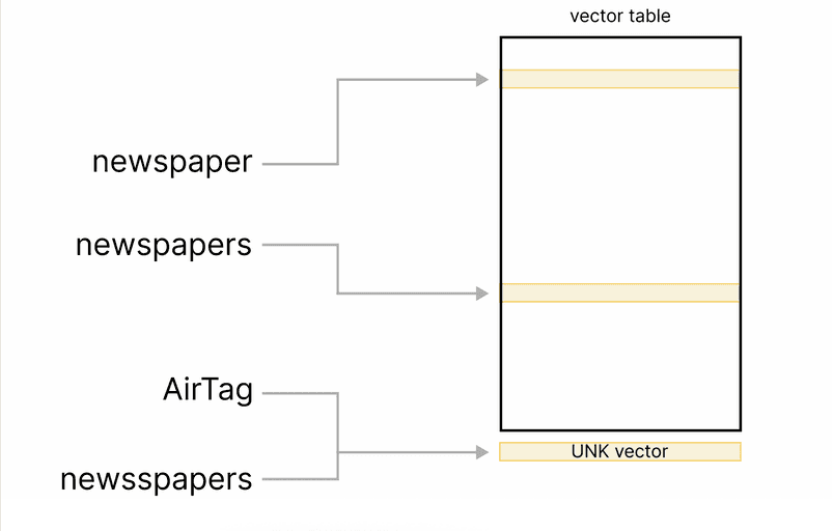
Performance of Compressed Models

Memory Usage

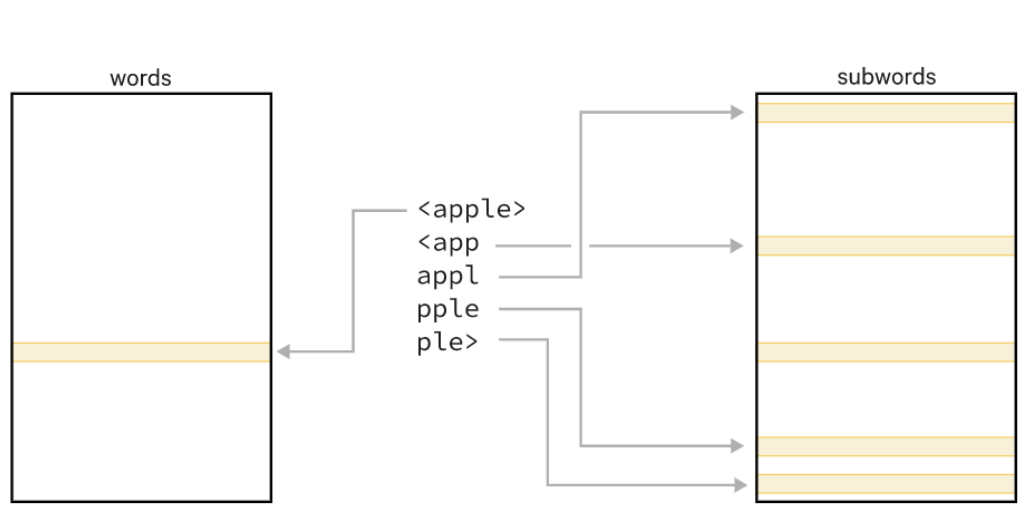


WORD EMBEDDINGS

Conventional



With subwords



https://explosion.ai/blog/floret-vectors

FASTTEXT EMBEDDINGS

Suitable for Slavonic (or other highly inflected languages)
Subwords – OOV problem practically disappears

Pretrained model for Czech by Facebook research (7GB):

- Wikipedia (179M tokens, 785k freq >= 5)
- CommonCrawl (13B tokens, vocab size = 8.7M)

CZECH FASTTEXT EMBEDDINGS 2022

Trained with the same parameters on:

- Wikipedia (179M → 218M, 785k → 863k)
- SYN v.9 (5.9B, vocabulary 3M)

Larger model (7GB → 10GB)

WORD EMBEDDING COMPRESSION METHODS

vocabulary — Remove rare tokens

dimensionality — Reduce vector dimensionality:

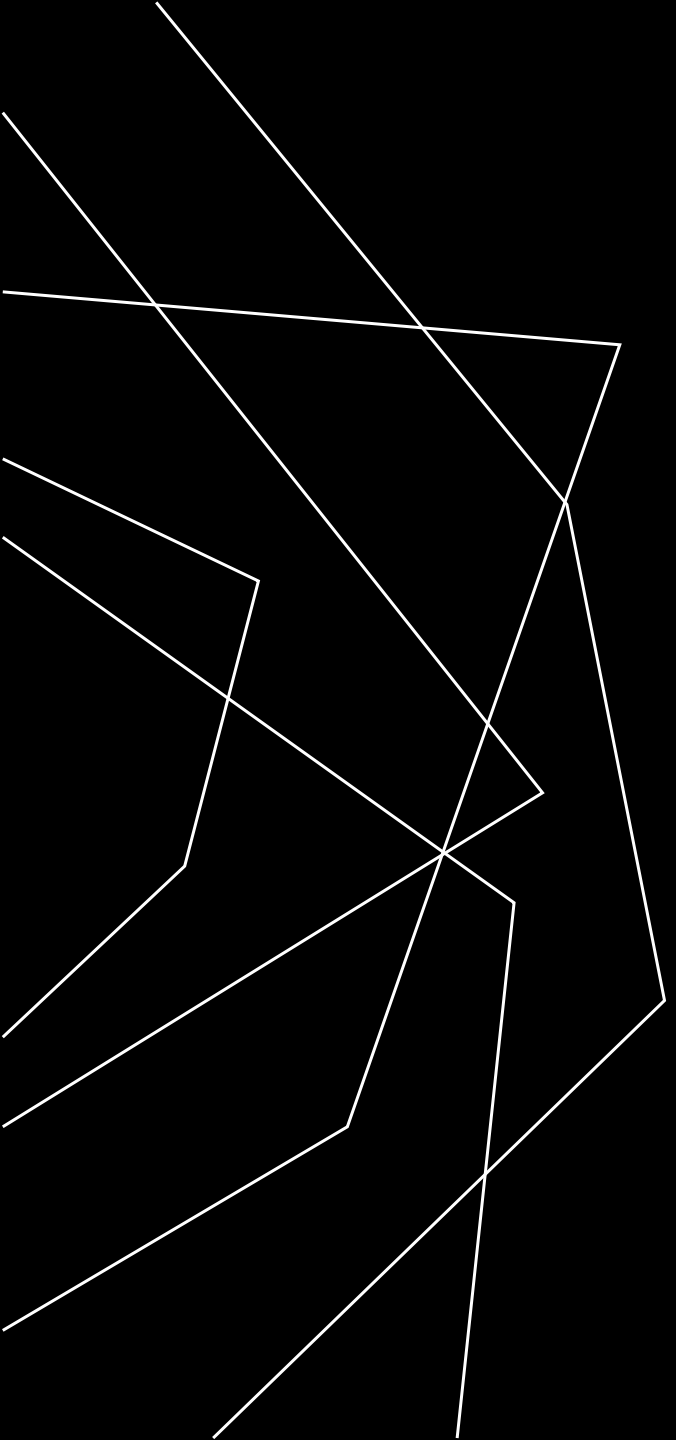
- Clever way = remove dimensions that do not affect the classification

pruning — Small weights set to zero

- Model becomes sparser

quantization — Weights “rounded” to nearest neighbors

- Model uses only a small number of different weights

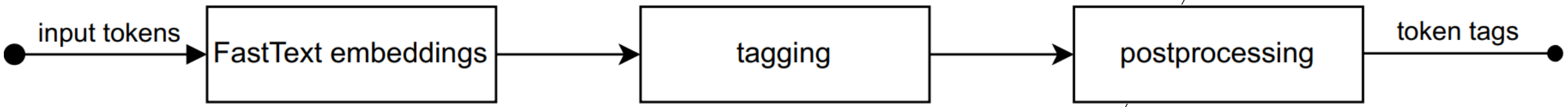


NEURAL TAGGER FOR CZECH

Mainská kočka mývalí

Top modelky

NEURAL TAGGER PIPELINE



Tagging as a multiclass classification task:

- Each token has 0-20 tag attributes (e.g. noun, plural)
- Some attributes are exclusive
- Some attribute combinations are invalid
  (e.g. preposition + past tense)

THE TAGGING PART

Neural network with two BiLSTM layers

Input width: 20 tokens (max. sentence length)

Tag attributes: 44 (simplified tagset)

Trained on 300k sentences from csTenTen17

Model size ~7MB

THE POSTPROCESSING PART

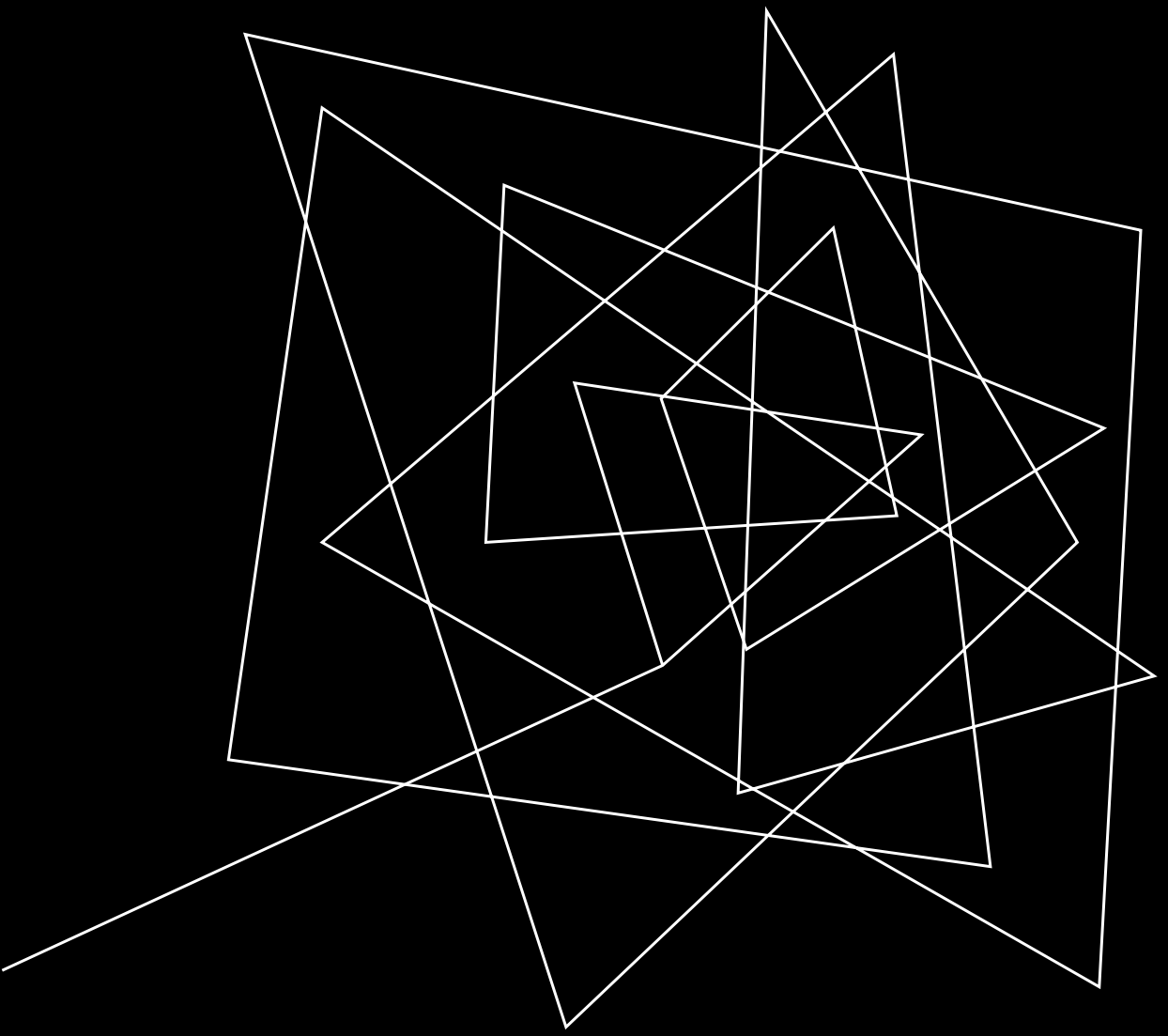
The tagger only predict tag attributes

No lemma

No attribute validity check

majka does the job (optional)

For majka OOV, a guesser is needed



COMPRESSED MODEL EVALUATION

Precision, recall, model size

TAGGING AND MAJKA POSTPROCESSING EXAMPLE

Top	modelky	chodí	po	molu
	k1gFnPc1	k5mIp3nP	k7c6	k1gInSc6
majka	k1gFnPc1	k5eAaImIp3nPrD	k7c6	k1gInSc6
	modelka	chodit	po	mol
mainská	kočka	mývalí		
k2gFnSc1d1	k1gFnSc1	k1gMc1		
majka	k1gFnSc1			
	kočka			

COMPRESSION METHOD, MODEL SIZE, AND PERFORMANCE (WITHOUT MAJKA)

Model name	Size	Precision	Recall	Exact matches
cc.cs.300.bin	**6.8GB**	**0.93**	**0.91**	**79.52**
cc.cs.300 prune freq	70MB	0.93	0.89	77.80
cc.cs.300 prune freq pq	14MB	0.93	0.89	77.12
cc.cs.300 quantize	426MB	0.93	0.90	78.88
cc.cs.300 svd	273MB	0.92	0.86	75.10
syn wiki.bin	**9.9GB**	**0.95**	**0.93**	**83.40**
syn wiki prune freq	70MB	0.94	0.92	82.33
syn wiki prune freq pq	14MB	0.94	0.92	82.07
syn wiki quantize	**587MB**	**0.95**	**0.93**	**83.11**
syn wiki svd	381MB	0.93	0.91	79.84

MAJKA IMPROVES THE RESULT

Quantized model without majka

P = 0.95

R = 0.93

Exact match = 83.11%

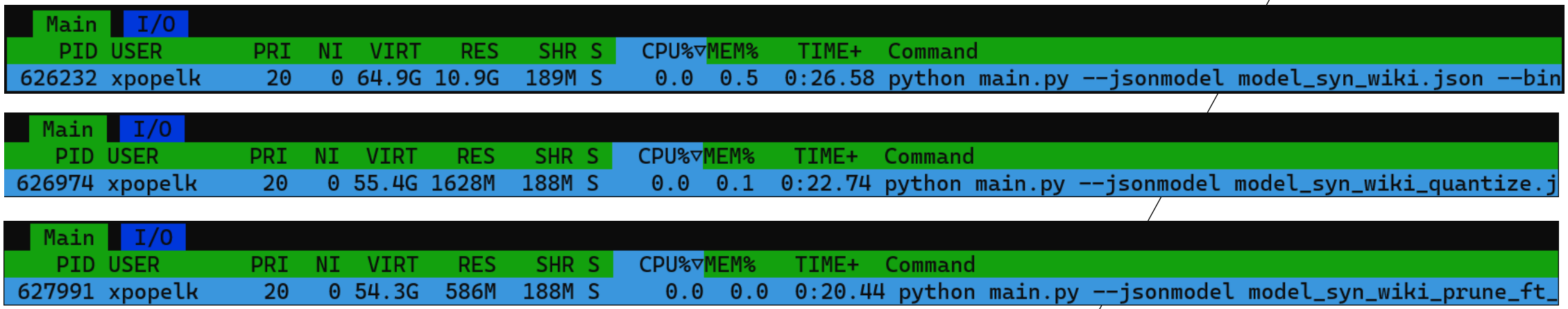
Quantized model with majka

P = 0.95

R = 0.93

Exact match = 90.53%

MEMORY USAGE BY THE SERVER



TODO

Process sentences > 20 tokens

Build stable version

- Check input format

Deploy

Add to Language Services
https://nlp.fi.muni.cz/languageservices/\

SUMMARY

Our own FastText embeddings outperform the Facebook FastText embeddings for Czech.

Compression does not affect tagger quality much.

It reduces the model size + memory usage by 1-3 orders.