Texterra @ ISP RAS

Transfer Learning for Morphological Tagging in Russian

Ivan Andrianov
Vladimir Mayorov

ISP RAS Open 2017

Morphological tagging task

- Assign a morphological tag to each word with respect to its context

	Мама	была	очень	рада
	Mom	*was*	*very*	*glad*
Part-of-speech	Noun	Verb	Adverb	Adjective
Case	Nominative	—	—	—
Number	Singular	Singular	—	Singular
Gender	Feminine	Feminine	—	Feminine
Animacy	Animated	—	—	—
Shortness	—	—	—	Short
Tense	—	Past	—	—
Mode	—	Indicative	—	—

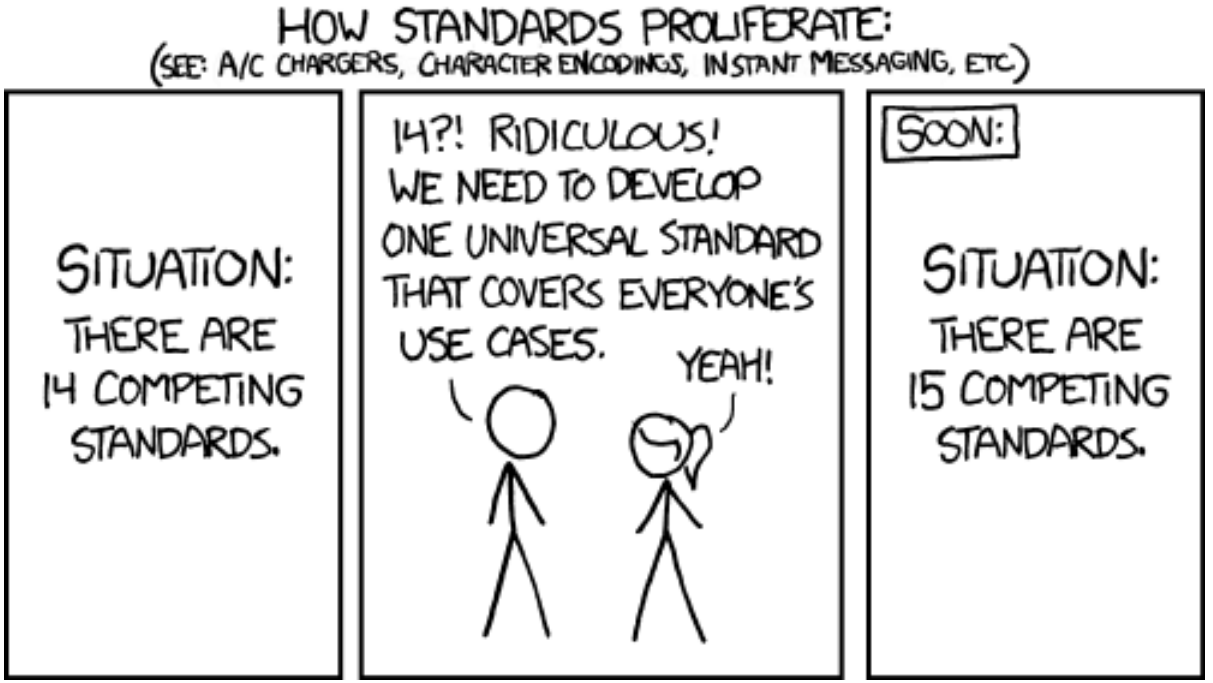
Morphologically annotated corpora for Russian before 2017

- There is no such thing as *Russian morphology*
- Linguists do not agree on, for example, how many cases do we have in Russian: 6 or 8
- Machine learning methods require unified tag sets to be properly trained on all data

	Tokens	PoS	Morph.
Ruscorpora	1.3 M	14	1321
SynTagRus[1]	1.1 M	12	459
SynTagRus UD[1]	1.0 M	17	700

MorphoRuEval-2017

- Shared task for Russian morphological tagging
- Introduced 4 new* corpora with unified tag sets
- Best participants employed the only one as joint training performed worse



Source: https://imgs.xkcd.com/comics/standards.png

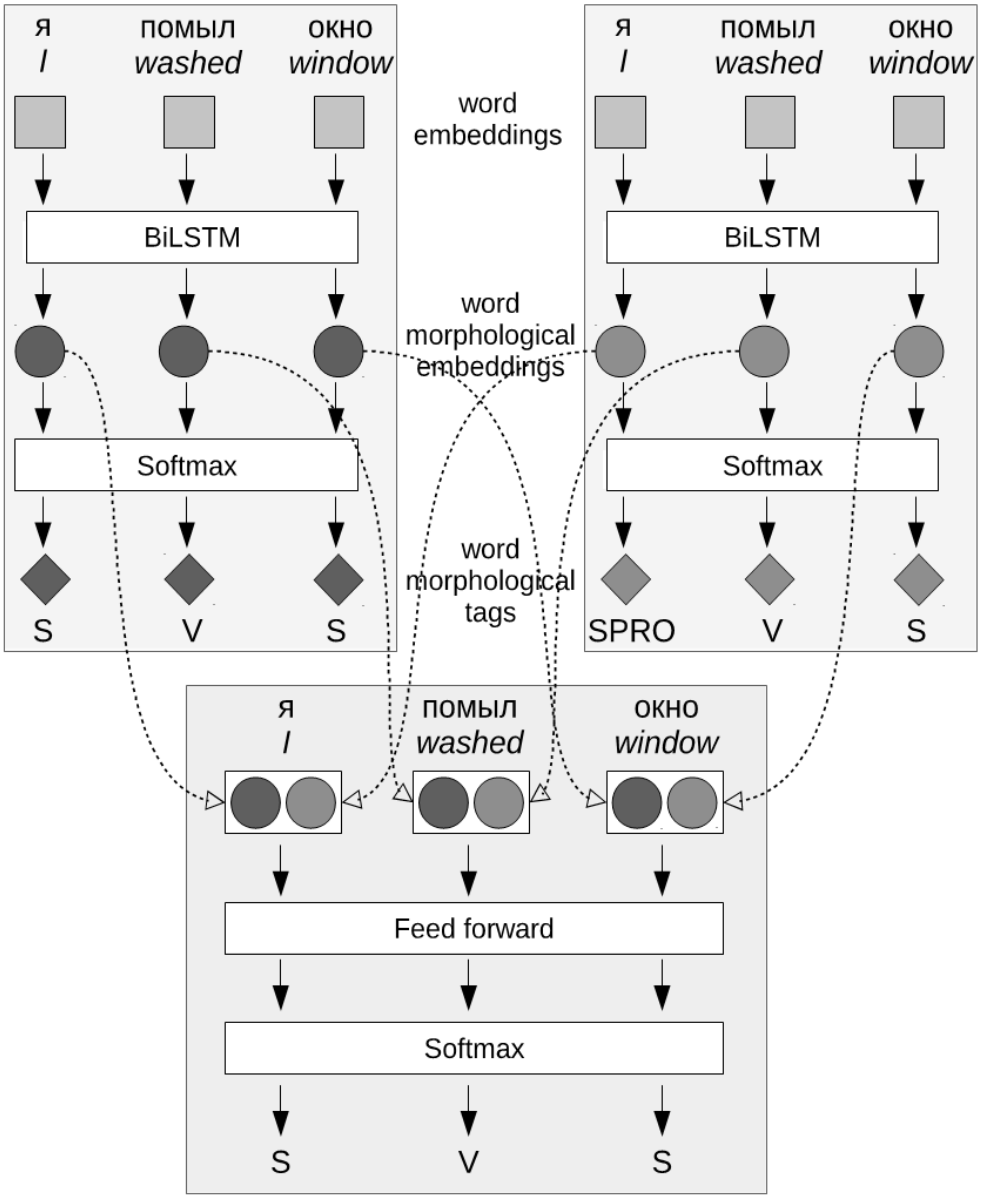
Morphologically annotated corpora for Russian after 2017

	Tokens	PoS	Morph.
Ruscorpora	1.3 M	14	1321
SynTagRus[1]	1.1 M	12	459
SynTagRus UD[1]	1.0 M	17	700
MorphoRuEval GICR	1.1 M	13	303
MorphoRuEval SynTagRus[1]	0.9 M	13	250
MorphoRuEval OpenCorpora	0.5 M	14	397*
MorphoRuEval RNC	1.35 M	15	2146*

Training against incompatible tag sets

- We have to have a separate classifier for each tag set but what about reusing features?
- Neural networks are known to be good at learning feature representations:
  - We can employ recurrent neural networks to construct feature representation of a word with respect to both its left and right contexts
  - To do this we additionally need a word embedding model: word2vec or fasttext

Ensemble transfer learning NN architecture



Experiments: «Classic» corpora

- Averaged measures of 5 random holdouts (90% / 10%) on target corpus

target	SynTagRus UD		SynTagRus		Ruscorpora	
subsid.	Ø	RNC	Ø	RNC	Ø	STR
$Full_{word}$	92.42	**93.35**	93.32	**93.96**	92.04	**92.78**
$Full_{sent}$	35.07	**38.97**	39.94	**43.68**	42.81	**45.22**
POS_{word}	97.37	**97.68**	97.68	**97.91**	97.38	**97.66**
POS_{sent}	65.50	**68.73**	68.81	**71.42**	70.32	**72.69**
POS F_1	86.08	**87.15**	86.04	**86.96**	92.99	**93.46**

Experiments: MorphoRuEval

- Testing on MorphoRuEval gold data

	News		Social media		Fiction		Altogether	
	Acc word	Acc sent	Acc word	Acc sent	Acc word	Acc sent	Acc word	Acc sent
GICR	92.20	47.21	89.86	54.58	90.77	48.48	90.97	50.76
-RNC	**94.35**	56.98	91.02	57.57	92.11	53.55	92.54	56.21
All	94.23	56.42	91.33	60.21	92.63	55.08	92.77	57.65
MSU-1	93.71	**64.80**	92.29	**65.85**	**94.16**	**65.23**	**93.39**	**65.29**
IQMEN	93.99	63.13	92.39	64.08	92.87	60.91	93.08	62.71
Sagteam	93.35	55.03	**92.42**	63.56	92.16	56.60	92.64	58.40

Conclusions

- Unification of morphological tag sets by hand is a labour-intensive and error-prone task
- Transfer learning improves quality consistently for all datasets by incorporating knowledge from subsidiary corpora
- fasttext word embedding model has a better sense of morphology than word2vec one thanks to its awareness of word character composition
- Results will be available at Texterra website

Thanks for your attention.
Any questions?