MorphoBabushka: Simple and Fast Baselines your Granny would use for Part-Of-Speech Tagging of Russian

Arefyev N. V. (nick.arefyev@gmail.com), Lomonosov Moscow State University, Moscow, Russia

Ermolaev P. A. (ermolaev.p.a@yandex.ru), Lomonosov Moscow State University, Moscow, Russia

MorphoRuEval: exercise for NLP students

POS-tagging: an example of

- multiclass classification (window-based) or
- sequence labeling (sentence-based) ...

Resources:

- 3 NLP students + their scientific supervisor
- 1 week (11 days after deadline was moved)

NN models to adapt:

- CharWNN (dos Santos et.al., 2014) @ Theano for NER (1 MSc)
- (Yoon Kim, 2014) @ Tensorflow for sentence classification (1 Msc)

  + conv over chars - conv over words

Baselines — bag of character n-grams repr. of each token +

- Sentence-based: CRF (1 BSc)
- Window-based: NB-SVM + other linear classifiers (me)

Results

Bag of character n-grams:

- NB-SVM is the best!
- Linear SVM is the 2nd, better than Logistic Regression, Multinomial NB, even Multilayer Perceptron (their scikit-learn impls!) using same input representations
- CRF lose (bad impl? cooked improperly?)

ConvNets are extremely slow to train

- 10 hours vs. 10 minutes (implemented by students?)

  => difficult to select hyperparameters

ConvNets could not beat best baselines (need better hyperparameters? use RNNs?)

Results

Bag of character n-grams:

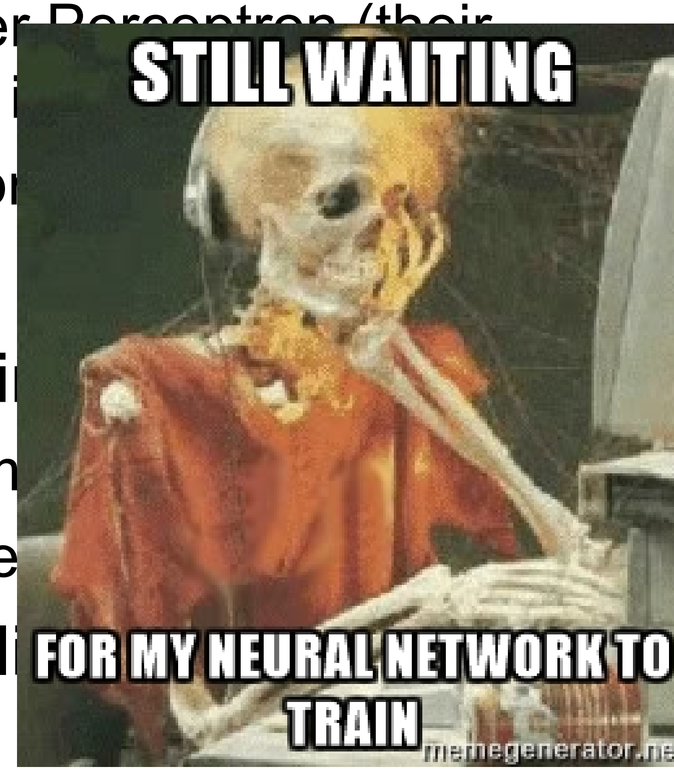
- NB-SVM is the best!
- Linear SVM is the 2nd, better than Logistic Regression, Multinomial NB, even Multilayer Perceptron (their scikit-learn impls!) using same i
- CRF lose (bad impl? cooked impr

ConvNets are extremely slow to trai

- 10 hours vs. 10 minutes (implem

  => difficult to select hyperparame

ConvNets could not beat best baseli hyperparameters? use RNNs?)



Datasets

Closed Shared Task

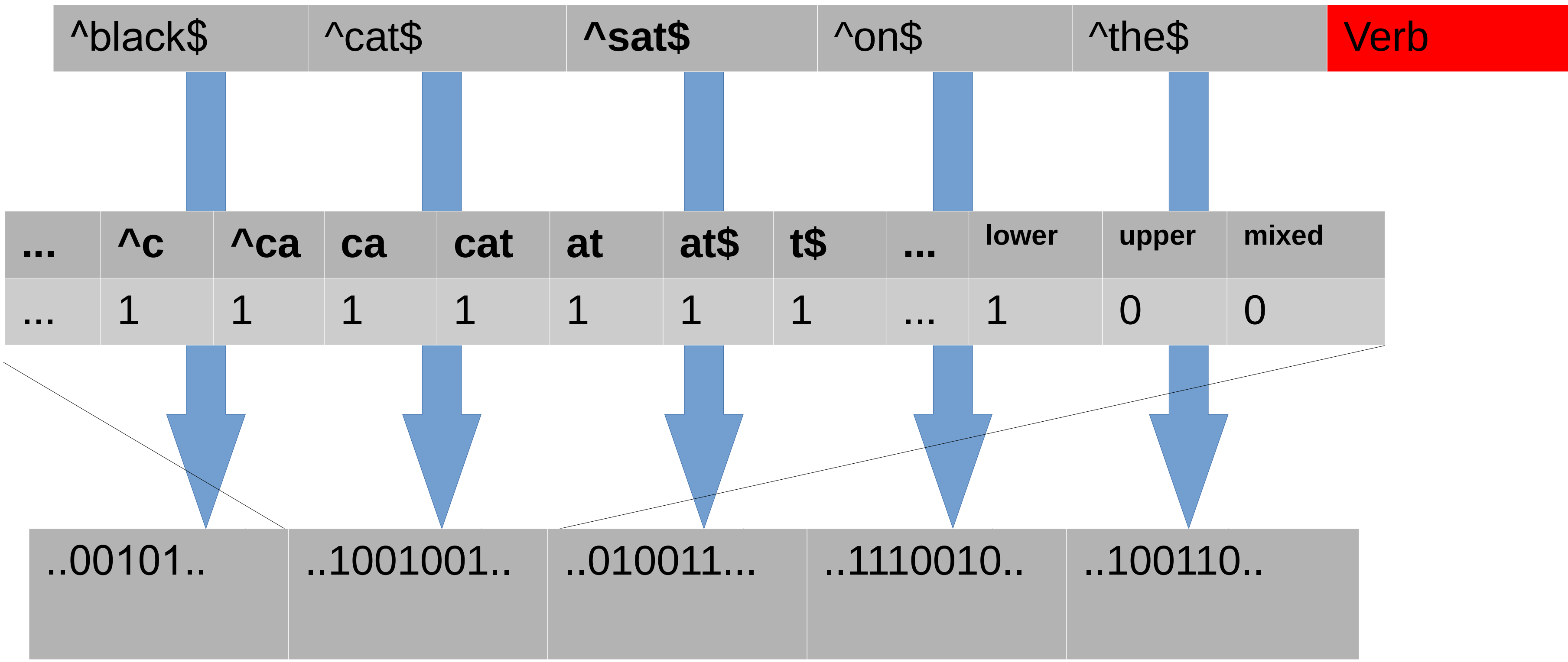
Gikrya, official train / test split:

- 62K / 20K sentences
- 815K / 270K tokens
- 100K — 500K tokens (windows) depending on grammatical category

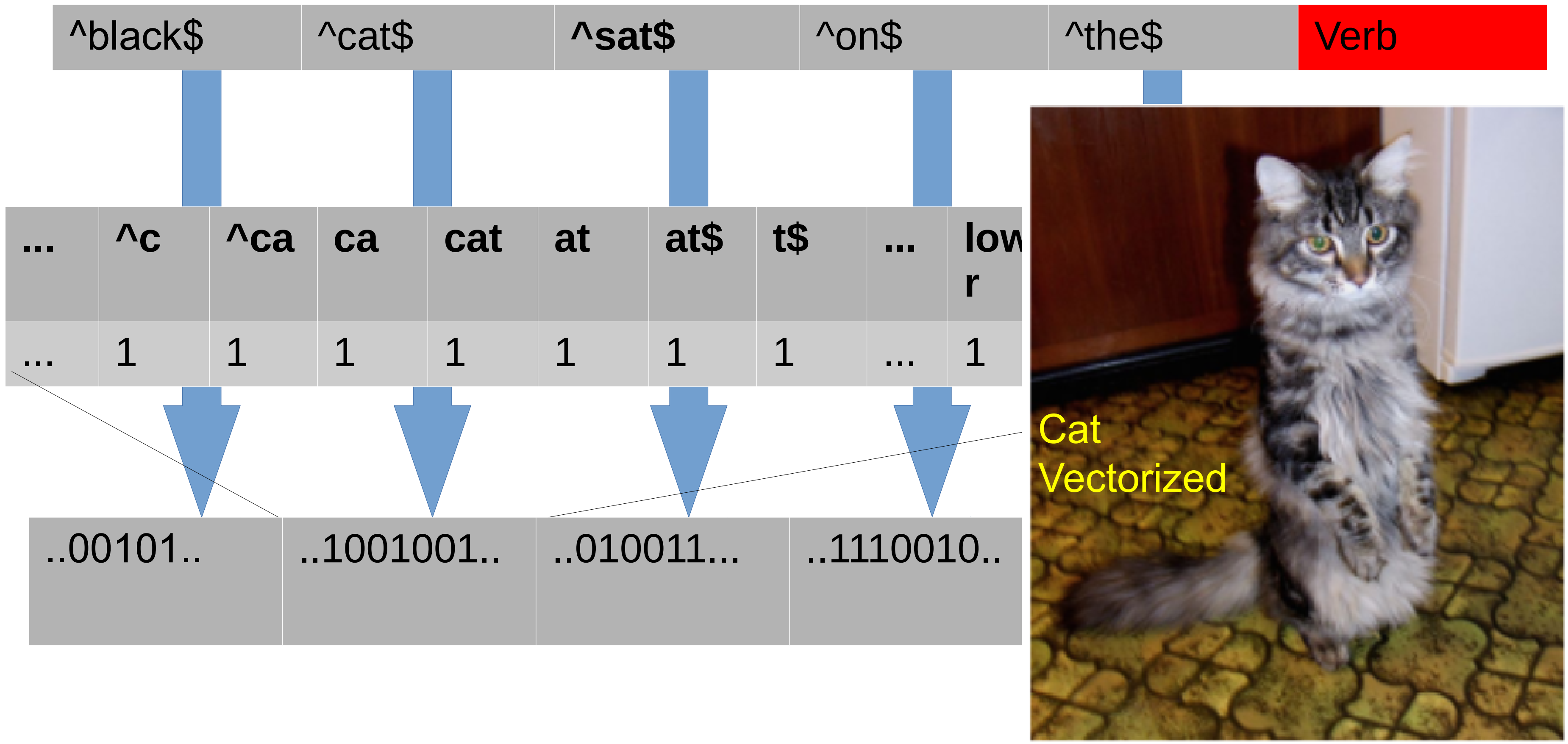
Did not use:

- Other labeled corpora
- Unlabeled corpora
- Dictionaries/word lists (including supplied by organizers determiners and pronouns lists)

Window Vectorization



Window Vectorization



NB-SVM classifier - Math

NB-scaling + linear SVM

$$r_i = log\left(\frac{p_i/||p||_1}{n_i/||n||_1}\right)$$

Proportion of i-th n-gram in positiv examples

Proportion of i-th n-gram in netgative examples

$$p_i = \alpha + \sum_k I\{y^{(k)} = +\}f_i^{(k)}$$

$$n_i = \alpha + \sum_k I\{y^{(k)} = -\}f_i^{(k)}$$

NB-scaling accounts for correlation between n-grams and classes (unlike Tf-idf scaling)

NB-SVM classifier - successes

Sentiment / topic classification, word n-grams:

- (Wang, Manning, 2012) ← year 1 before w2v

Baselines and Bigrams: Simple, Good Sentiment and Topic Classification

Method	RT-s	MPQA	CR	Subj.
MNB-uni	77.9	85.3	79.8	**92.6**
MNB-bi	**79.0**	**86.3**	80.0	**93.6**
SVM-uni	76.2	86.1	79.0	90.8
SVM-bi	77.7	**86.7**	80.8	91.7
NBSVM-uni	**78.1**	85.3	80.5	92.4
NBSVM-bi	**79.4**	**86.3**	**81.8**	**93.2**
RAE	76.8	85.7	–	–
RAE-pretrain	77.7	**86.4**	–	–
Voting-w/Rev.	63.1	81.7	74.2	–
Rule	62.9	81.8	74.3	–
BoF-noDic.	75.7	81.8	79.3	–
BoF-w/Rev.	76.4	84.1	**81.4**	–
Tree-CRF	77.3	86.1	**81.4**	–
BoWSVM	–	–	–	90.0

Our results	RT-2k	IMDB	Subj.
MNB-uni	83.45	83.55	**92.58**
MNB-bi	85.85	86.59	**93.56**
SVM-uni	86.25	86.95	90.84
SVM-bi	87.40	**89.16**	91.74
NBSVM-uni	87.80	88.29	92.40
NBSVM-bi	**89.45**	**91.22**	**93.18**
BoW (bnc)	85.45	87.8	87.77
BoW (bΔt'c)	85.8	88.23	85.65
LDA	66.7	67.42	66.65
Full+BoW	87.85	88.33	88.45
Full+Unlab'd+BoW	**88.9**	88.89	88.13
BoWSVM	87.15	–	90.00
Valence Shifter	86.2	–	–
tf.Δidf	88.1	–	–
Appr. Taxonomy	**90.20**	–	–
WRRBM	–	87.42	–
WRRBM + BoW(bnc)	–	**89.23**	–

Method	AthR		XGraph		BbCrypt	
MNB-uni	85.0		90.0		**99.3**	
MNB-bi	**85.1**	+0.1	**91.2**	+1.2	99.4	+0.1
SVM-uni	82.6		85.1		98.3	
SVM-bi	83.7	+1.1	86.2	+0.9	97.7	−0.5
NBSVM-uni	**87.9**		**91.2**		99.7	
NBSVM-bi	**87.7**	−0.2	**90.7**	−0.5	99.5	−0.2
ActiveSVM	–		90		99	
DiscLDA	83		–		–	

NB-SVM classifier - successes

Sentiment classification, word n-grams:

– (Mesnil et.al., 2015) ← year 2 after w2v

Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews

Single Methods	Accuracy
N-gram	86.5%
RNN-LM	86.6%
Sentence Vectors	88.73%
NB-SVM Trigram	91.87%

Ensemble	Accuracy
RNN-LM + NB SVM Trigram	92.13%
RNN-LM + Sentence Vectors	90.4%
Sentence Vectors + NB-SVM Trigrams	92.39%
All	**92.57%**
State of the art	91.22%

NB-SVM classifier - successes

POS-tagging, character n-grams:

– This work

Table 1. Accuracy on POS-tagging. NB-SVM (no padding) doesn't add special symbols (^ and $) to the token. NB–SVM (no caps) doesn't use capitalization features

accuracy	model
0.93	Memory baseline
0.97	CRF
0.979	NB-SVM (no padding)
0.98	Tf-idf + linear SVM
0.981	linear SVM
0.983	NB-SVM (no caps)
0.983	NB-SVM

NB-SVM — our implementation

Probably, NB-SVM is the best linear classifier for NLP!

We brought it to Scikit-Learn!

Our NB-SVM impl.

- Scikit-learn compatible
- Extention: scaling schemes (binarize or scale features before NB-scaling), separately for train/test sets
  - Best scaling scheme depends on dataset
  - Crossvalidate to select best scaling scheme
  - Random search to select scaling scheme and regularization simultaneously

https://github.com/nvanva/MorphoBabushka

sklearn_ext/nbsvm.py

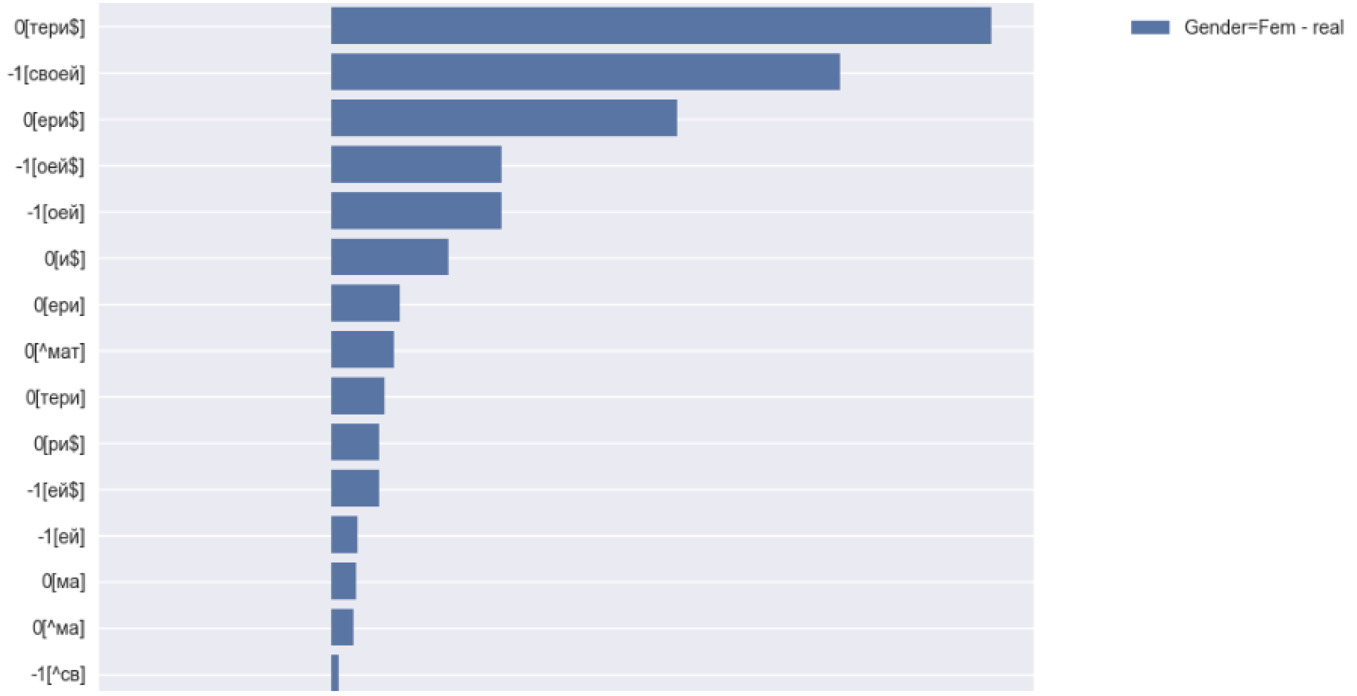
Real example

Document from dev set

	-2	-1	0	1	2	-2_cap	-1_cap	0_cap	1_cap	2_cap	y_true
215998	^o$	^своей$	^матери$	^,$	^в$	0	0	0	0	0	Gender=Fem

25 most and least weighted ngrams

- 0[тери$]
- -1[своей]
- 0[ери$]
- -1[оей$]
- -1[оей]
- 0[и$]
- 0[ери]
- 0[^мат]
- 0[тери]
- 0[ри$]
- -1[ей$]
- -1[ей]
- 0[ма]
- 0[^ма]



Window / n-grams sizes

Need character n-grams with n>3

(for bag of word n-grams even n=3 helps very little)

win=1 is bad

win>3 does not help

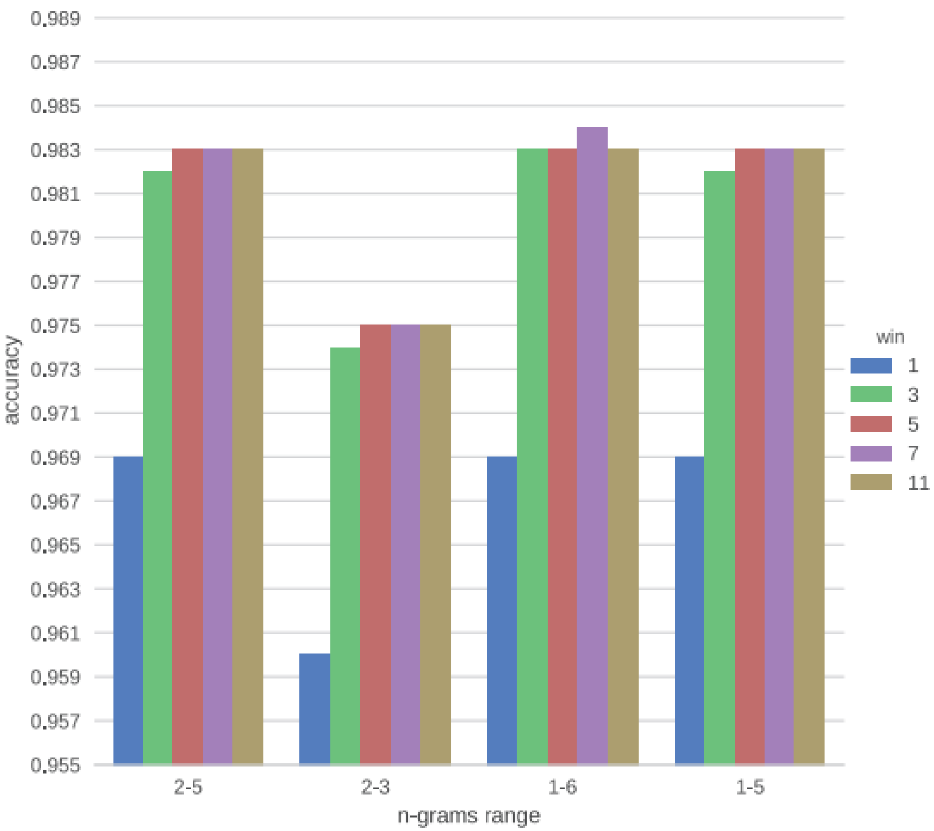


Figure 2. Accuracy of NB-SVM on POS-tagging w.r.t. window and n-grams sizes

Grammatical categories

What about Case, Number, Gender, etc.?

1. Single output => >200 classes

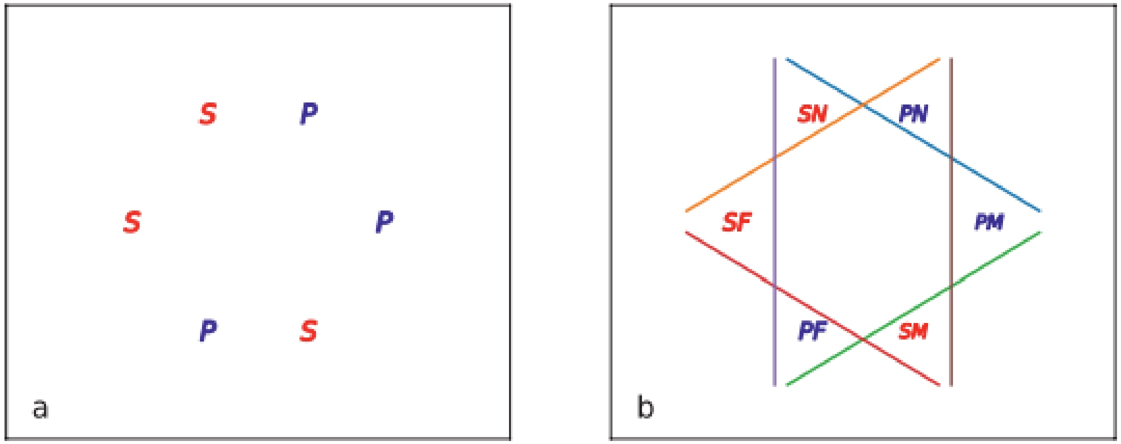
    `Pos=VERB+Gender=Neut+Mood=Ind+Number=Sing+Tense=Past+VerbForm=Fin`

2. Multiple output, 1 per grammatical category

3. Group outputs

    `Case=Ins+Number=Sing`

Grouping



Grouping helps!

grouping	number of outputs	accuracy
—	10	0.922
Gender+Number+Case, VerbForm+Mood+Tense	6	0.926
Gender+Number	9	0.923
Number+Case	9	0.928
VerbForm+Mood+Tense	8	0.922

Final results

Grouping was done after deadline :(

classificator	dev accuracy (per token)	test accuracy (per-token/per-sentence)
NB-SVM	0.921	0.901 / 0.481
CRF	0.913	0.892 / 0.456
Memory baseline	0.742	0.724 / 0.138
NB-SVM (grouping—Number+Case)	0.928	—

				Новости				Вконтакте			
Team name	team ID	дорожка	Номер лучшей попытки	точность по меткам	точность по предложениям	лемматизация, точность по словоформам	Лемматизация, точность по предложениям	точность по меткам	точность по предложениям	лемматизация, точность по словоформам	Лемматизация, точность по предложениям
МГУ-1, Алекс	C	закрытая	2	93,71	64,8			92,29	65,85		
IQMEN	O	закрытая	1	93,99	63,13	92,96	56,42	92,39	64,08	91,69	61,09
Sagteam	H	закрытая	2	93,35	55,03	81,6	17,04	92,42	63,56	82,8	35,92
Аспект, НИИ	A	закрытая	2	93,83	61,45	93,01	54,19	91,49	61,44	90,97	60,21
Morphobabush	M	закрытая	2	90,52	44,41			89,55	51,41		
Pullenti Pos Tag	G	закрытая	4	89,73	39,66	89,04	37,71	89,17	54,58	88,65	52,64
	B	закрытая	6	90,79	43,58			88,96	52,29		
	N	закрытая	4	91,53	49,16	87,01	25,7	88,44	48,59	83,67	34,51
	K	закрытая	4	90,36	45,53	89,23	40,22	88,39	52,11	87,34	48,94
	F	закрытая	2	90,43	36,87	89,61	33,52	86,72	44,72	85,81	41,9
	I	закрытая	2	88,66	29,89			84,29	41,73		
	L	закрытая	2	75,88	2,79			70,13	14,61		

Errors

~50% - errors in Case

~25% - errors in Pos

	accuracy	error number	error rate	support
Pos	0.983	4,537	0.017	270,264
Number	0.984	2,298	0.016	142,411
Case	0.927	8,117	0.073	110,967
Gender	0.979	2,262	0.021	107,544
VerbForm	0.999	31	0.001	39,083
Mood	0.998	64	0.002	30,170
Tense	1.000	0	0.000	31,227
Variant	1.000	0	0.000	3,810
NumForm	1.000	0	0.000	925
Degree	0.999	60	0.001	40,608

Errors

POS:

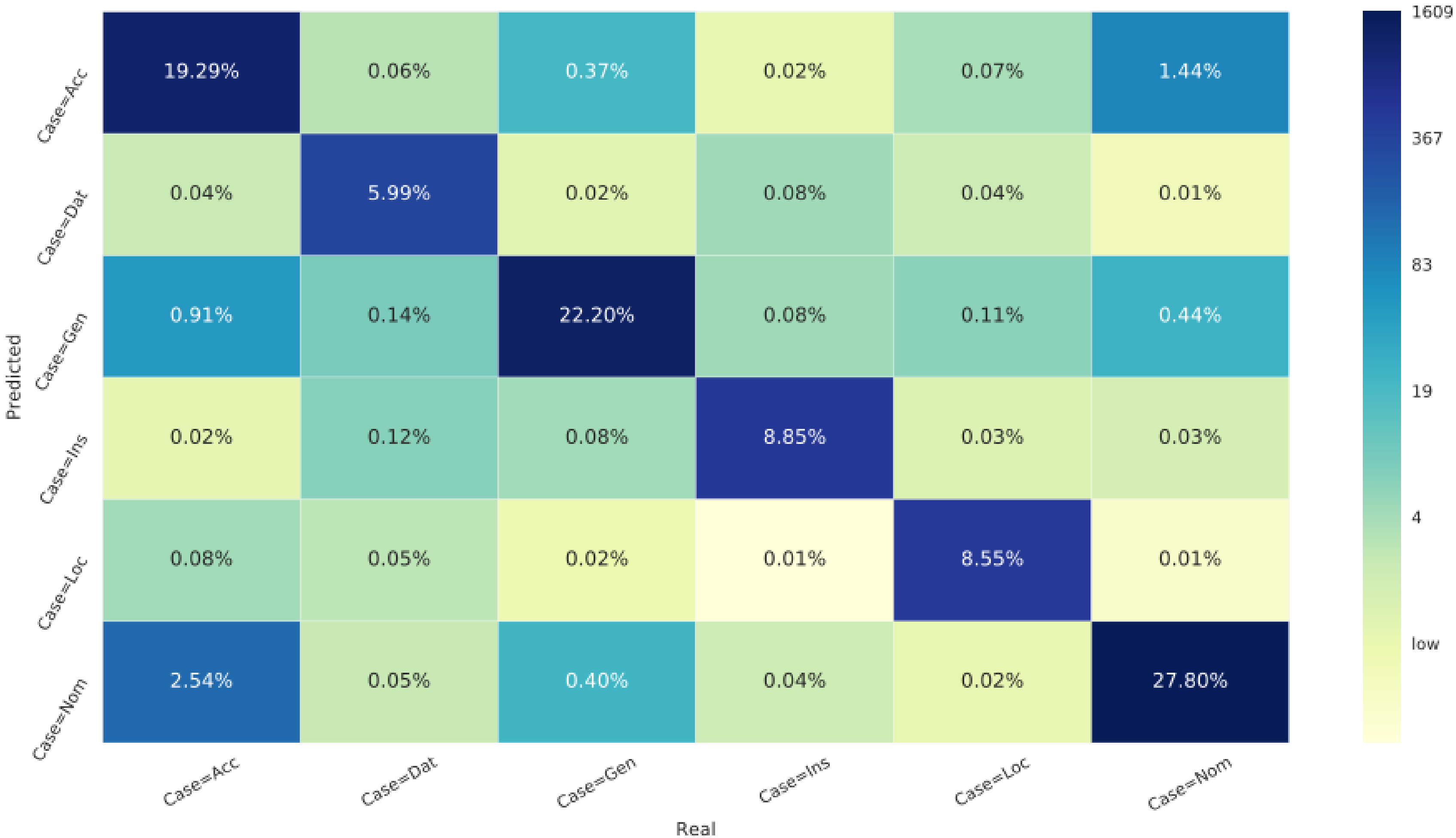
– INTJ, CONJ

Case:

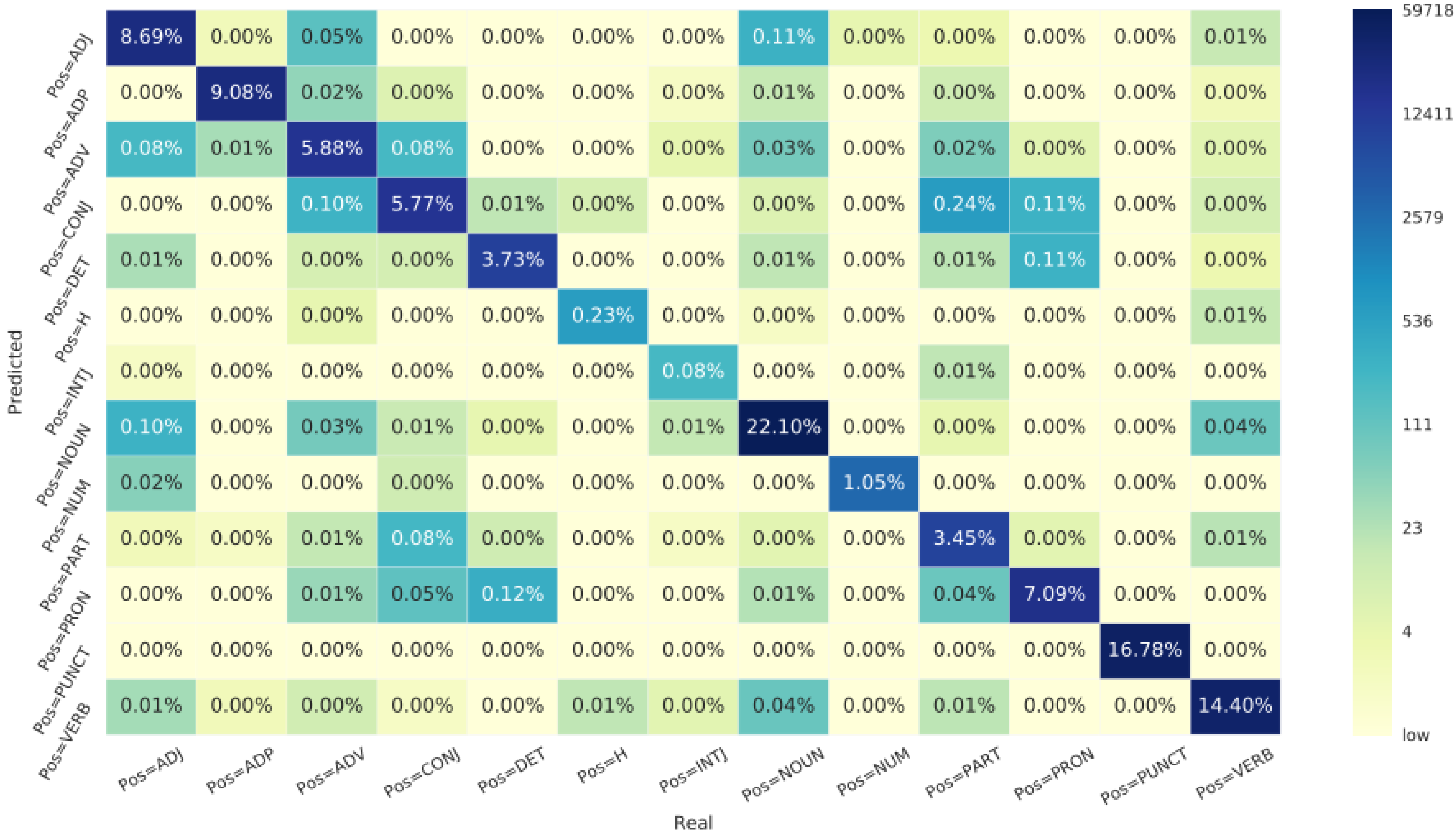
– Nom, Acc

	precision	recall	f1-score	support
Pos=ADJ	0.98	0.97	0.98	24,113
Pos=ADP	1.00	1.00	1.00	24,573
Pos=ADV	0.96	0.96	0.96	16,498
Pos=CONJ	0.93	0.96	0.94	16,211
Pos=DET	0.96	0.96	0.96	10,442
Pos=H	0.96	0.96	0.96	651
Pos=INTJ	0.91	0.87	0.89	257
Pos=NOUN	0.99	0.99	0.99	60,271
Pos=NUM	0.98	1.00	0.99	2,855
Pos=PART	0.97	0.91	0.94	10,208
Pos=PRON	0.97	0.97	0.97	19,742
Pos=PUNCT	1.00	1.00	1.00	45,360
Pos=VERB	1.00	1.00	1.00	39,083
Number=Plur	0.98	0.96	0.97	38,009
Number=Sing	0.99	0.99	0.99	104,402
Case=Acc	0.91	0.84	0.87	25,389
Case=Dat	0.97	0.93	0.95	7,112
Case=Gen	0.93	0.96	0.95	25,615
Case=Ins	0.97	0.97	0.97	10,070
Case=Loc	0.98	0.97	0.98	9,791
Case=Nom	0.90	0.94	0.92	32,990
Gender=Fem	0.99	0.98	0.98	35,574
Gender=Masc	0.97	0.98	0.98	48,054
Gender=Neut	0.98	0.97	0.98	23,916

Confusion matrix: Case



Confusion matrix: Pos



Thank you

Questions?