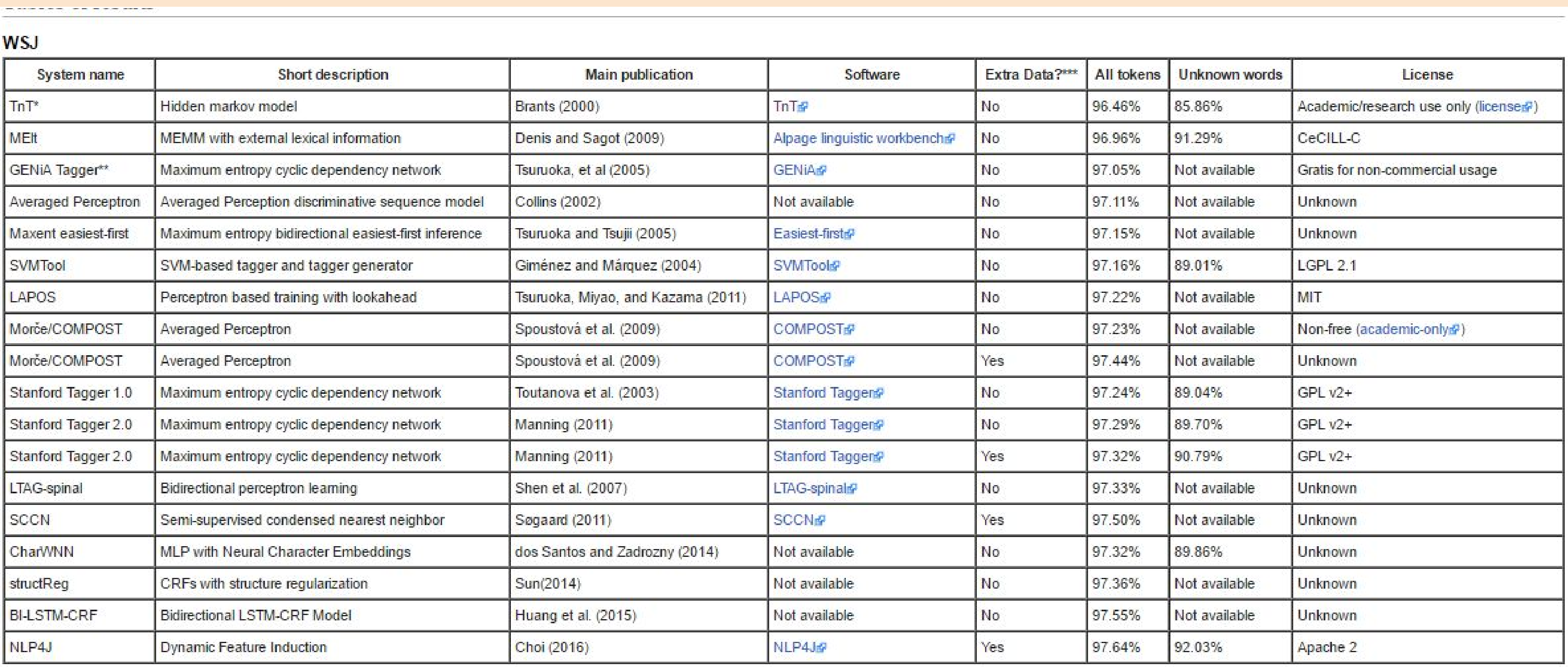
Sorokin A.A., Shavrina T.O., Lyashevskaya O.V., Bocharov V., Alexeeva S., Droganova K., Fenogenova A., Granovsky D.

MorphoRuEval-2017: an Evaluation Track for the Automatic Morphological Analysis Methods for Russian

01.06.2017 International Conference Dialogue 2017, Moscow, Russia

POS tagging for English



WSJ

System name	Short description	Main publication	Software	Extra Data?***	All tokens	Unknown words	License
TnT*	Hidden markov model	Brants (2000)	TnT	No	96.46%	85.86%	Academic/research use only (license)
MElt	MEMM with external lexical information	Denis and Sagot (2009)	Alpage linguistic workbench	No	96.96%	91.29%	CeCILL-C
GENiA Tagger**	Maximum entropy cyclic dependency network	Tsuruoka, et al (2005)	GENiA	No	97.05%	Not available	Gratis for non-commercial usage
Averaged Perceptron	Averaged Perception discriminative sequence model	Collins (2002)	Not available	No	97.11%	Not available	Unknown
Maxent easiest-first	Maximum entropy bidirectional easiest-first inference	Tsuruoka and Tsujii (2005)	Easiest-first	No	97.15%	Not available	Unknown
SVMTool	SVM-based tagger and tagger generator	Giménez and Márquez (2004)	SVMTool	No	97.16%	89.01%	LGPL 2.1
LAPOS	Perceptron based training with lookahead	Tsuruoka, Miyao, and Kazama (2011)	LAPOS	No	97.22%	Not available	MIT
Morče/COMPOST	Averaged Perceptron	Spoustová et al. (2009)	COMPOST	No	97.23%	Not available	Non-free (academic-only)
Morče/COMPOST	Averaged Perceptron	Spoustová et al. (2009)	COMPOST	Yes	97.44%	Not available	Unknown
Stanford Tagger 1.0	Maximum entropy cyclic dependency network	Toutanova et al. (2003)	Stanford Tagger	No	97.24%	89.04%	GPL v2+
Stanford Tagger 2.0	Maximum entropy cyclic dependency network	Manning (2011)	Stanford Tagger	No	97.29%	89.70%	GPL v2+
Stanford Tagger 2.0	Maximum entropy cyclic dependency network	Manning (2011)	Stanford Tagger	Yes	97.32%	90.79%	GPL v2+
LTAG-spinal	Bidirectional perceptron learning	Shen et al. (2007)	LTAG-spinal	No	97.33%	Not available	Unknown
SCCN	Semi-supervised condensed nearest neighbor	Søgaard (2011)	SCCN	Yes	97.50%	Not available	Unknown
CharWNN	MLP with Neural Character Embeddings	dos Santos and Zadrozny (2014)	Not available	No	97.32%	89.86%	Unknown
structReg	CRFs with structure regularization	Sun(2014)	Not available	No	97.36%	Not available	Unknown
BI-LSTM-CRF	Bidirectional LSTM-CRF Model	Huang et al. (2015)	Not available	No	97.55%	Not available	Unknown
NLP4J	Dynamic Feature Induction	Choi (2016)	NLP4J	Yes	97.64%	92.03%	Apache 2

POS tagging for English and Russian

- POS-tagging for English:
  - Relatively simple morphology.
  - Established training corpus (WSJ Penn Treebank).
  - Multiple approaches (HMM, CRF, dependency networks, neural network methods).
  - High baseline.
- POS-tagging for Russian:
  - Large number of tags.
  - Ubiqitous homonymy.
  - Long-distance dependencies.
  - Fine-grained categories, complex interaction between them.

POS tagging for Russian

- POS-tagging algorithm for Russian:
  - No reference corpora.
  - No comparison of different algorithms.
  - Problems with baseline approaches.
- Possible algorithms:
  - HMM cannot extract all the information.
  - CRF require too much memory, cannot handle long distance dependencies.
  - Discriminative-based approaches not tested, possible large number of features.
  - Neural network approaches not tested.
- Our goals:
  - Provide a reference corpus.
  - Compare different algorithms.
  - Determine directions for future work.

Russian NLP Evaluation Initiative

Previous Russian Morphology forum:

Ru-Eval 2010 * state-of-the-art, mostly rule-based taggers, test dataset (2k words)

- POS & Lemmatization: 13 answers
- Morphology: 12 answers
- Rare words: 8 answers

- Disambiguation (POS, Lemma): 7 answers

Soft evaluation: 94.5 - 95% accuracy

2017 Tracks

1. **Closed track:** the participants are allowed to train their models only on provided data.

-for research groups and student teams

-own dictionaries allowed

2. **Open track:** track members are allowed to bring any data for learning

-for enterprise participants

Full morphological tags are evaluated and also (optionally) lemmatization.

Tagset

Universal Dependencies 1.4 and 2.0

Parts of Speech:
noun (NOUN), proper name (PROPN), adjective (ADJ), pronoun (PRON) numeral (NUM), verb (including auxiliary, VERB), adverb (ADV), determinant (DET), conjunction (CONJ), preposition (ADP), particle (PART), interjection (INTJ).

Also marked: punctuation marks (PUNCT), non-word tokens (X), parenthesys (H).

Omitted: SYM (symbol) and AUX (auxiliary verb).

Tagset

Case	nominative - Nom, genitive - Gen, dative - Dat, accusative - Acc, locative – Loc, instrumental - Ins
Gender	masculine - Masc, feminine - Fem, neuter - Neut
Number	singular - Sing, plural - Plur
Animacy	animate - Anim, inanimate - Inan
Tense	past - Past, present or future - Notpast
Person	first – 1, second – 2, third - 3
VerbForm	infinitive - Inf, finite - Fin, gerund - Conv (participles are treated as ADJ)
Mood	indicative – Ind, imperative - Imp
Variant	short form – Brev (no tag for long form)
Degree	positive or superlative - Pos, comparable - Cmp
NumForm	numeric token – Digit (if the token is written in alphabetic form, no mark is placed).

Tagset problems and solutions

Categorical mismatches in different data sources:

- no Aspect tags on GICR & Syntagrus data : Tense (past, present, future) → Tense (past, notpast)
- PROPN and NOUN are equally evaluated
- There is a special “H” tag for parenthetic constructions
- Participles and ordinal numerals are considered adjectives, gerunds - part of a verb paradigm
- Predicatives are considered short forms of adjectives, with an exception for *“нет”*, which is a verb

Some categories listed explicitly: Determiner (UD), Conjunctions, Particles, Prepositions, Parenthesis, Pronouns.

Conj, Part, Prep, Int, H, X and some adverbs (*как, пока, так, когда*), homonymic to them, are not taken into account during evaluation. GICR data was proposed as a standard if any other differences occurred.

Data

For both tracks we provide the following training data:

annotated data:

1) RNC Open: a manually disambiguated subcorpus of the Russian National Corpus - 1.2 million words (fiction, news, nonfiction, spoken, blog)

2) GICR corpus with the resolved homonymy - 1 million words

3) OpenCorpora.org data - 0.4 million tokens

4) UD SynTagRus - 0.9 million tokens (fiction, news)

And also **plain text data**: 1) LiveJournal (from GICR) 30 million words 2) Facebook, Twitter, VKontakte - 30 million words 3) Librusec - 300 million words

All data available at https://github.com/dialogue-evaluation/morphoRuEval-2017

Data

Test set:

1. news texts (Lenta.ru)
2. fiction (Russian Magazine Hall, magazines.russ.ru)
3. social networks (vk.com) – from unpublished part of GICR materials for MorphoRuEval (other data resources were previously published).

600-900 thousand tokens for each segment.

Gold standard

3 different segments from GICR for testing – Lenta.ru, fictions (Russian Magazine Hall), VK

7000 tokens each

Baseline:

TreeTagger trained on annotated MorphoRuEval data

best - 79% accuracy per tag depending on the sources of testing and training data

best - 26% accuracy per sentence.

Evaluation and metrics

Four **metrics** for ranking:

- Percentage of correct tags and tag-lemma pairs (in case the system outputs lemmas).
- Percentage of correctly labeled sentences both by tags and by tag-lemma pairs.
- All metrics are calculated for three subtasks and for the whole dataset.
- Sentence accuracy on the entire dataset used for ranking.

Reason: subtle differences in tag accuracy become significant for sentence accuracy.

We also used the following conventions:

1) Both PROPN and NOUN labels for proper nouns is correct. The same holds for SCONJ and CONJ with respect to conjunctions.

2) capitalization is not significant for lemmatization.

3) *е* and *ё* are not distinguished.

Evaluated categories

Evaluated POS and grammemes:

- Nouns (gender, number, case).
- Adjectives (gender, number, case, degree, brevity).
- Verbs (mood, tense, person, gender, number).
- Determiners (gender, number, case).
- Pronouns (gender, number, case, person).
- Numerals (gender, case, graphic form)
- Adverbs (degree)

Not evaluated:

- Conjunctions,
- Prepositions,
- Particles,
- Parentheses,
- Punctuation.

Other information provided by participants is not taken into account

Competition results

Top 6 teams (of 15 participants)

Team name	Track	Tags	Sents	Lemma	Lemma sents
ABBYY	Open	97,11	83,68	96,91	82,13
MSU-1	Closed	93,39	65,29	-	-
IQMEN	Closed	93,08	62,71	92,22	58,21
Sagteam	Closed	92,64	58,4	80,73	25,01
Aspect	Closed	92,57	61,01	91,81	56,49
Morphobabushka	Closed	90,07	48,1	-	-

Competition algorithms

Several types of algorithms for morpho tagging:

- Neural networks (ABBYY (clear winner), Sagteam, Aspect)
- Classification-based (IQMEN, Morphobabushka).
- Reranking-based (MSU).

Lemmatization algorithms: usually a conversion pattern is guessed using the same features as for the tag itself.

Most of the participants train on GICR subset of the training data.

ABBYY team additionally train on Wiktionary corpus annotated by Compreno parser.

Top-ranked algorithms

- The clear winner **ABBYY team**
- LSTM network as main classier.
- Several layers in the network (up to 10).
- Two types of features on input layer:
  - Grammatical and suffix features extracted using Compreno parser.
  - Pre-trained word embeddings fine-tuned on the training set.

Top-ranked algorithms

- Second team: **MSU**, winner on the closed track. An attempt to build linguistically oriented system.
- HMM as basic classier generating hypotheses.
- Initial training data extended with transitivity for verbs and case for prepositions.
- Hypotheses are reranked using high-level features.
- Examples of features:
    - Number of coordinated adjective-noun groups.
    - Number of coordinated preposition-noun groups.
    - Number of nominative nouns coordinated with verbs.
    - Number of transitive verbs having a direct object.
- Learning algorithm: generate hypotheses for the sentences in the training set and train a linear classifier on the differences between top hypothesis and the others.
- Decision algorithm: select the hypothesis with the highest score according to the classifier.

Top-ranked algorithms

- **Aspect team**: bidirectional LSTM.
  - Separate character, flexion and stem embedding on the input layer.
  - Embeddings for stem and flexions trained on LibRuSec corpora.
  - Several dense layers in the network.
- **Sag**: convolutional neural network.
  - Character level embedding on the input layer for individual words.
  - Several layers in the network.

Top-ranked algorithms

- **IQMEN team**: window-based classification approach.
    - Each word is tagged in isolation using the features from surrounding words and the word itself.
    - Word features: word prex and sux up to 4 characters.
    - Left context features: POS and tag features (case, gender, number) for all the words in 7-word window.
    - Right context features: ambiguity classes for POS and tag features.
    - SVM with hash kernel as the final classifier.
- **Morphobabushka** team:
    - No dictionary used, word is tagged using only features.
    - Word features: character ngrams. Features are extracted from the word itself and its neighbours.
    - A classifier outputs the labels for tag features (e.g. case) or combinations of features (number + gender).
- NB-SVM classier

Questions for future work

- Main problems for tagging morphologically rich languages:
  - Abundancy of morphological features and classes.
  - Long distance dependencies.
- Ways to overcome feature abundancy:
  - More powerful classifiers (neural networks).
  - Modification of more traditional classifiers in window-based approaches.
- Ways to deal with long-distance dependencies.
  - Convolutional layers in CNN, long-short memory in LSTM.
  - Global features in reranking.
  - Global features in classification (not used?).
- Role of training corpora and dictionaries?
  - Character-based approaches allow to work without dictionaries.
  - Neural networks require large corpora (though unannotated).
- Global approaches on the top of neural network model (neural network provides local correctness, global features stand for global constraints).

Conclusion

Comparing to previous evaluation of morphological parsers for Russian language, current systems show significant improvement. Indeed, the top-ranked of the [Lyashevskaya et al.,2010] competition achieved 97% result only for POS-tagging, while the winner of current competition showed the same result for entire grammatical tags. The top-system result is comparable with results for other inflective languages with free word order and rich inflective morphology, such (95.75% for Czech in [Strakova, 2013]).

Shared task on morphological tagging showed fruitful results in several important aspects:

- An original data set collected from different corpora which was annotated in a single format consistent with UD guidelines was prepared and presented;
- the comprehensive guidelines for testing procedure and evaluation were created.
- The comparison of different parsing strategies showed that neural network approach is state-of-the-art method for morphological parsing of Russian.
- A dataset for future improvement of morphological parsers, comprising texts from different sources, was created.