Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2017"

Moscow, May 31—June 3, 2017

MORPHORUEVAL-2017: AN EVALUATION TRACK FOR THE AUTOMATIC MORPHOLOGICAL ANALYSIS METHODS FOR RUSSIAN

Sorokin A. (alexey.sorokin@list.ru)[1,2,6],
Shavrina T. (rybolos@gmail.com)[4,6],
Lyashevskaya O. (olesar@yandex.ru)[4,8],
Bocharov V. (victor.bocharov@gmail.com)[3,5],
Alexeeva S. (sv.bichineva@gmail.com)[3],
Droganova K. (kira.droganova@gmail.com)[4,9],
Fenogenova A. (alenka_s_ph@mail.ru)[4,7],
Granovsky D. (dima.granovsky@gmail.com)[3]

[1]Lomonosov Moscow State University, [2]MIPT,
[3]OpenCorpora.org, [4]National Research University Higher
School of Economics, [5]Yandex, [6]GICR, [7]RDI KVANT,
[8]Vinogradov Institute of the Russian Language RAS,
[9]Charles University

MorphoRuEval-2017 is an evaluation campaign designed to stimulate the development of the automatic morphological processing technologies for Russian, both for normative texts (news, fiction, nonfiction) and those of less formal nature (blogs and other social media). This article compares the methods participants used to solve the task of morphological analysis. It also discusses the problem of unification of various existing training collections for Russian language.

Key words: shared task, morphological tagging, morphological parsing, parsers for Russian, universal dependencies, automatic morphological analysis, POS tagging, disambiguation, taggers

MORPHORUEVAL-2017: ОЦЕНКА МЕТОДОВ АВТОМАТИЧЕСКОГО МОРФОЛОГИЧЕСКОГО АНАЛИЗА НА МАТЕРИАЛЕ РУССКОГО ЯЗЫКА

Сорокин А. (alexey.sorokin@list.ru)[1,2,6],
Шаврина Т. (rybolos@gmail.com)[4,6],
Ляшевская О. (olesar@yandex.ru)[4,8],
Бочаров В. (victor.bocharov@gmail.com)[3,5],
Алексеева С. (sv.bichineva@gmail.com)[3],
Дроганова К. (kira.droganova@gmail.com)[4,9],
Феногенова А. (alenka_s_ph@mail.ru)[4,7],
Грановский Д. (dima.granovsky@gmail.com)[3]

[1]МГУ им. М.В.Ломоносова, [2]МФТИ, [3]OpenCorpora.org,
[4]Национальный Исследовательский университет
«Высшая школа экономики», [5]ООО «Яндекс»,
[6]ГИКРЯ, [7]НИИ КВАНТ, [8]Институт русского языка
им. В. В. Виноградова РАН, [9]Карлов Университет (Прага)

MorphoRuEval-2017 — соревнование по морфологической разметке, призванное стимулировать развитие технологий морфологической обработки текстов на русском языке, в особенности текстов из сети Интернет, как нормативных (новости, литературные тексты), так и менее формального характера (блоги и другие социальные медиа). Данная статья посвящена сравнению методов, использованных командами-участниками соревнования, а также проблемам унификации различных существующих обучающих коллекций для русского языка.

Ключевые слова: соревнование по морфологическому анализу, частеречная разметка, автоматическая морфологическая разметка, алгоритмы морфологической разметки для русского языка, снятие омонимии

1. Introduction

Russian morphology has a long history of extensive research, both theoretical and practical. While theoretical science faces a wide range of problems concerning distinction of parts of speech and classification of grammatical categories [Sichinava 2011], practice of NLP usually finds temporary solutions, which occur less or more acceptable and convenient. There are already several morphological tagsets for Russian, all of them derived from different approaches such as MSD for Russian, AOT tags, OpenCorpora.org tags, Russian Positional Tagset, Natural Language Compilator tagset, etc. Generally, these tagsets are not convertible into each other without loss of information.

There already exist several solutions dealing with this problem[1], yet there is no single candidate to use as reference tagset, e. g. in evaluation tracks. The only open and internationally acknowledged tagset—MSD—is overly fine-grained for the purpose of tagset unification. The morphological data standard for shared task should be 1) concise 2) compatible with international shared task results 3) suitable for rapid and consistent annotation by a human annotator 4) suitable for computer parsing with high accuracy 5) easily comprehended and used by a non-linguist (last 3—Manning Laws [Nivre, 2016])). A plausible solution of the problem is a new standard of multilingual morphological and syntactic tagging—Universal Dependencies[2] (UD) [Nivre et al. 2016]. UD initiative has developed 70 treebanks for 50 languages with cross-linguistically consistent annotation and recoverability of the original raw texts, and apparently the UD standard is becoming the main annotation paradigm for many languages. Continuing the tradition of independent evaluation of the methods used in Russian language resources and linguistic tools [Lyashevskaya et al. 2010; Toldova et al. 2012], we designed this evaluation track of Russian morphological analysis methods in order to inspire the development of the morphological taggers. For that purpose we presented the original training set which was annotated in a single format consistent with UD guidelines.

2. Evaluation tracks

Within the competition framework, we relied heavily on the experience of the previous morphological forum of Dialogue Evaluation [Lyashevskaya et al. 2010]. However, we decided to refuse organizing the track without disambiguation: participants should give only one answer for each token even if it requires disambiguation, which is the question of interest in our case. Another innovation in such campaigns for Russian is dividing the competition on the basis of model training conditions: an open track and a closed one.

1. **Closed track:** the participants are allowed to train their models only on provided data. Mostly, it is convenient for research groups and student teams who do not have large data collections. To verify the results, participants of this track are required to make their code publicly available on github, both for organizers and other participating teams. This track was intended for comparison of various tagging algorithms. Since no dictionary in competition format is available, the participants of the closed track might use their own dictionaries as well after converting them to competition format.
2. **Open track:** track members are allowed to bring any data for learning (this regulation is more appropriate for enterprise participants presenting their products).

For both tracks we provide the following evaluation (see 5):

- POS-tagging;
- tagging of the categories of interest;
- lemmatization.

[1] https://github.com/kmike/russian-tagsets

[2] http://universaldependencies.org/

The key goal of the competition was to test comparative strength of different tagging methods in two setups: a closed one, which evaluates the ability of the algorithm to learn from limited data, and the open, which allows the tagger to use any possible source of data. Since the 90-s, the state-of-the-art in morphological tagging were variants of Hidden Markov Models, where the probability of the next tag was calculated either using ngram models, as in TnT Tagger [Brants, 2000] or by the means of decision trees, as in Tree Tagger [Schmid, 1995]. For English they were beaten by conditional random fields [Sha, Pereira, 2002] and dependency networks [Toutanova, 2002], however, for the languages with developed inflected morphology their advantage is not so clear, if any, since the number of features grows too fast with the order of model. Therefore when using CRF for tagging, for example, Czech or German, one has to make decoding more complicated [Muller, 2013]. Recent advances of neural networks in POS tagging [Huang, 2015] makes them a perspective candidate.

There is no clear benchmark for morphological tagging for Russian. The previous competition organizers [Lyashevskaya et al. 2010] give no analysis of results; a recent work of [Dereza et al., 2016] shows that HMM-based approach combined with decision trees realized in Tree Tagger are substantially ahead of others, however they give no error analysis and their results are not reproducible. Two main features of Russian are free word order and regular homonymy between different forms of the same word (e. g., nominative and accusative of inanimate noun) which cannot be resolved by immediate context of the word. Hence the applicability of standard HMM or CRF approaches is limited since they cannot capture, for example, the coordination between the noun and the verb in the sentence in case these words are divided by more than 2 words. Therefore it is not clear, whether the usage of more powerful methods of machine learning or more linguistically-oriented algorithms is more beneficial. One of the goals of current competition was to investigate this dichotomy.

3. Participants

The competition was joined 15 research groups from 7 universities and research institutes (MSU, NSU, MIPT, NRU HSE, ISPRAS, NRCKI, MIEM) and 5 companies (Abbyy, OnPositive, Pullenti, Samsung R&D Institute Moscow, IQMEN) and also 3 independent researchers.

The competition resulted in 11 teams providing their materials for the closed track, and 5 teams for the open one. 1 participant have succeeded to take part in both tracks (with slight improvement on open track). About half of the teams have presented their results with lemmatization, while 7 have provided only their tagging.

4. Collecting the training data

It was decided to collect as much as possible of the annotated data in a single format for training, and additionally, to provide a sufficient number of plain texts of different genres, for participants to obtain lexical frequencies, information on compatibility and syntactic behavior, vector embeddings, etc. In total, MorphoRuEval-2017 provided the following resources:

plain texts:

1) LiveJournal (from GICR) 30 million words
2) Facebook, Twitter, VKontakte—30 million words[3]
3) Librusec—300 million words

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—400 thousand tokens
4) UD SynTagRus—900 thousand tokens (fiction, news)

To unify the representation of the marked data, the conll-u format was chosen, as the most common and convenient, and for the unification of morphological tags—the format of the Universal Dependencies (further UD) 2.0 (with some specifications, see below). Resulting text collections are now available under CC BY-NC-SA 3.0 license.

4.1. Unification of the morphological tagset in annotated data

Remaining within the UD framework, we nevertheless decided to abandon some of the agreements adopted in this format to facilitate the procedure for unifying the training set. As part of the unification, we did not set the task of reducing the whole tokenization to a single variant, and we specified some complex tokens existing in GICR and UD Syntagrus data (they received the label "H").

We omitted two POS tags SYM (symbol) and AUX (auxiliary verb), keeping in out collection the following part-of-speech categories: 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 on the data are marked punctuation marks (PUNCT) and non-word tokens (X).

The following categories are annotated:

1. Noun: gender, number, case, animate
2. Proper name: gender, number, case
3. Adjective: gender, number, case, brevity of form, degree of comparison
4. Pronoun: gender, number, case, person
5. Numeral: gender, case, graphic form
6. Verb: mood, person, tense, number, gender
7. Adverb: degree of comparison
8. Determinant: gender, number, case
9. Conjunction, preposition, particle, parenthesis, interjection, other: none

[3] We have collected posts and comments from random users and political posts for recent 5 years, fuzzy deduplication has been done to decrease the effect of popular and spam messages.

Table 1. Annotated categories for different parts of speech

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
Mood	indicative—Ind, imperative—Imp
Variant	short form—Brev (no mark for complete form)
Degree	positive or superlative—Pos, comparable—Cmp
NumForm	numeric token—Digit (if the token is written in alphabetic form, no mark is placed).

In order to increase the annotation agreement in the collections converted from different sources, the following decisions were made (most of them follow the guidelines of UD SynTagRus corpus):

1) DET is a closed class which includes 30 pronouns used primarily in the attributive position.
2) Predicative words. Modal words such as *можно* 'can', *нельзя* 'cannot' are considered as adverbs. The word *нет* 'no, not' is considered as verb. The predicative words homonymous to the short neuter forms of adjectives are coded as adjectives. Therefore, short adjectives always form a part of the predicate, while adverbs do not, which can be checked semi-automatically at least in sufficient fraction of cases. This solution was accepted to facilitate automatic verification and unification of different annotated corpora since they follow different disambiguation standards and even these standards often are not realized consistently. Moreover, even in the simplest cases the border between different categories is rather vague. Our final solution coincides with UD SynTagRus guidelines after joining together short adjectives and predicatives.
3) The lemma of the verb is its infinitive form in a particular aspect (perfective or imperfective). The gerund forms constitute a part of the verb paradigm. Since the voice category was excluded, verbs ending with reflexive verb suffix *–ся* also had *–ся* in their infinitive form (the infinitive of *пишется* in *книга пишется писателем* is *писаться*, not *писать*).
4) The participles are treated as adjectives and their lemma is the Nominative masculine singular form. This was done to avoid border cases between adjectives and participles. Therefore voice category is irrelevant both for participles and other types of verbs and it was excluded from competition evaluation
5) The ordinal numerals are considered as adjectives.
6) The tense forms of the verb are divided into Past and Notpast (present or future). Aspect is not evaluated to avoid problems with biaspectual verbs.

7) The analytic (multi-word) forms of verbs, adjectives, and adverbs are not coded. For example, the analytic future tense form is annotated as two separate tokens: the future form of the verb *быть* 'to be' and infinitive.
8) SCONJ and CONJ are embraced by a single category CONJ.

A number of categories received the status of "not rated": they may be present or not in the output of the system under evaluation:

- animacy (nouns, pronouns);
- aspect, voice, and transitivity (verbs);
- pos-tags of the prepositions, conjunctions, particles, interjections, and X (others).

Several adverbs (*как* '*how*', *пока* 'while, yet'*, так* 'so', *когда* 'when') homonymic to conjunctions were also not rated since their annotation was controversial in different training corpora and even inside the same training corpus.

These guidelines differ a bit from those accepted in Universal Dependencies version of SynTagRus. This was done in order to simplify verification of morphological tags and their unification across corpora. Nevertheless, some inconsistency is still present. The list below summarizes the most significant differences.

RNC Open:

1) in the CONLL-u format, an extra column is provided with additional tags for typos, non-standard inflectional forms, aspect, voice, transitivity, NameType categories, etc.

GICR (the same conventions hold for the test set of the competition)

1) PROPN is tagged as NOUN.
2) A number of multi-token parenthetics as well as some other multi-word expressions (marked as H) is preserved. These multi-token constructions could also appear in the test set.

OpenCorpora:

1) Homonymy between the comparative forms of adjectives and adverbs is always resolved as the forms of adjectives (due to the agreements in the OpenCorpora dictionary)
2) the verb aspect is tagged
3) the list of possible multitoken constructions slightly differs from UD SynTagRus and GICR.

UD SynTagRus:

1) PROPN is tagged as NOUN
2) A number of multi-token expressions (marked as H) is preserved.

Since both the test set and the part of the training set used by most of the competitiors is the GICR subcorpora, we describe in more details its annotation pipeline. Initially it was automatically processed by ABBYY Compreno parser[4] providing a high-quality automatic annotation. One of the benefits of this parser is extensive usage of semantic information which helps to resolve one of the most difficult types of homonymy

[4] https://www.abbyy.com/ru-ru/isearch/compreno/

in Russian morphology, the one between accusative and nominative cases. However, as every automatic annotation, it suffers from several problems of other type. What is even more important, annotation standards and morphological system of ABBYY Compreno differ significantly from the one of UD. For example, the system always treats *это* as a demonstrative pronoun, while in UD standard and its competition dialect it was considered as a pronoun when it serves as a subject *это было трудно* 'it was difficult' and as a determiner when it is an attribute *это решение было трудным* 'this solution was difficult'. The same problem holds for the word *все* (*все пришли вовремя* 'all came on time' vs *все мои друзья пришли вовремя* ('all my friends came on time'). These ambiguities are important for the quality of annotation since pronouns are very frequent. We checked during conversion whether a particular instance of such pronouns is a undoubtful attribute (it is followed by a noun in the same case, gender and number) or a subject (for example, it is followed by a corresponding form of auxiliary verb *быть*). Analogous constraints were applied to verify and correct annotation of adverbs (for example, a potential adverb appearing between subject and verb is an actual adverb *он легко ответил* 'he easily answered') and other frequent ambiguities.

5. Testing procedure

Competitors obtain a tokenized sample as a test set. They should assign a morphological tag and (optionally) lemma to all the words in the test sample. However, only the grammemes described in Section 4 are evaluated, the presence/absence of other categories does not affect the results of evaluation. It also does not matter, which label is assigned to the words whose parts of speech are not rated, such as conjunctions, prepositions etc.

The participants should strictly follow the requirements below:

1) POS and categories labels should be taken from https://github.com/dialogue-evaluation/morphoRuEval-2017/blob/master/morphostandard
2) The tokenization of the test set is preserved. A participant should tag all the sentences in the test sample and all the words in each sentence.
3) the unique text IDs are preserved (but ignored by tagging)

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.

5.1. Metrics

We evaluated participants performance on three test sets of different origin, News texts (Lenta.ru), fiction (Russian Magazine Hall, magazines.russ.ru) and social networks (vk.com). For each of the segments, two metrics were calculated: the percentage of correctly parsed words and the percentage of sentences whose entire parse was correct. If the participant provided lemmas, both tagging and full (lemma+tag) accuracy were evaluated, otherwise only the tag accuracy was considered. We also calculated average metrics across all three segments. For final ranking overall sentence accuracy was used since usually a correct parse of the whole sentence.

5.2. Baseline

In the review [Dereza etc., 2016] authors evaluated several taggers on the material of 6 million Russian National Disambiguated Corpus (mainly literary texts), the highest accuracy of 96,94% on POS tags and of 92,56% on the whole tagset was achieved by TreeTagger [Schmid, 1995]. This is a HMM-based tagger, which uses a binary decision tree to estimate transition probabilities. TreeTagger is also capable to tag the unknown words using a suffix/prefix lexicon. For current shared task TreeTagger was chosen as a baseline system.

We have carried on five baseline experiments (results are presented in the Table 2):

1. On the material of GICR: 75% training set, 25% test set.
2. Trained on the data of GICR: 75% training set; tested on the data of Syntagrus 25% test set.
3. On the material of Syntagrus: 75% training set, 25% test set.
4. Trained on GICR 75% training set, Syntagrus 75% training set and 1 million RNC and Opencorpora.org dataset. Tested on GICR 25% test set.
5. Trained on GICR 75% training set, Syntagrus 75% training set and 1 million RNC and Opencorpora.org data set. Tested on Syntagrus 25% test set.

Table 2. Evaluation of baseline algorithms for different training settings

Experiment	Tags	accuracy per tag	number of correct tags	accuracy per sentence	number of correct sentences
Baseline (1)	POS tag	79.49%	136,372 from 171,550	26.25%	5,456 from 20,787
	Full tag	76.54%	131,309 from 171,550	21.14%	4,394 from 20,787
Baseline (2)	POS tag	73.46%	107,846 from 146,817	9.93%	1,244 from 12,529
	Full tag	68.44%	100,482 from 146,817	6.35%	795 from 12,529
Baseline (3)	POS tag	79.19%	116,265 from 146,817	17.02%	2,132 from 12,529
	Full tag	75.43%	110,749 from 146,817	11.87%	1,487 from 12,529
Baseline (4)	POS tag	73.89%	126,759 from 171,550	23.89%	4,967 from 20,787
	Full tag	71.15%	122,054 from 171,550	18.51%	3,848 from 20,787
Baseline (5)	POS tag	72.10%	105,854 from 146,817	14.89%	1,866 from 12,529
	Full tag	69.71%	102,346 from 14,6817	11.76%	1,473 from 12,529

5.3. Golden Standard

We have provided 3 different segments from GICR for testing, all not published before, 7000 tokens each. These are 568 sentences from VKontakte, from News (Lenta ru, 353 sentences), and from modern literature, Russian Magazine Hall (394 sentences).

These materials were tagged morphologically within the framework of GICR pipeline [Selegey et al. 2016], then converted from MSD to UD 1.4 and carefully checked automatically and manually, with paying special attention to consistency of annotation, format specifications and systematic errors of automatic tagging (case homonymy, short form adjectives and adverbs, etc.). As a side result we discovered, that no existent automatic or semi-automatic procedure guarantees the quality of morphological analysis sufficient to be a "Gold standard" for parsers testing and manual verification and correction is a necessary postprocessing step. Golden standard sentences were randomly shuffled in a tokenized set of 600–900 thousand tokens for each segment.

All participant results and scripts for their comparison with golden standard are now available online[5].

6. Team results and methods

One of the goals of current competition was to compare different approaches to morphological tagging. The clear winner of the competition is ABBYY team which participated in the open track. In the closed track slightly better than others was the team of MSU, however three other teams are less than 1% behind in terms of tag accuracy, the gap for sentence accuracy is more significant.

The top-ranked participant algorithms fall in two camps. The first utilizes the power of neural networks to uncover hidden relationships between words in the sentence. It includes the winner (ABBYY) and Sagteam and Aspect from the closed track. The second group tries to use linguistic information using more complex features. It contains MSU and IQMEN teams (top 2 on closed track).

ABBYY team uses two-layer bidirectional neural network with several additional layers as a learning method. Each word is characterized by 250-dimensional embedding and additional morphological and graphic features. The model was pre-trained on additional Wikipedia corpus and these parameters were further optimized on GICR data from the training set. Wikipedia corpus was pretagged using ABBYY Compreno parser.

MSU team used an HMM classifier as a baseline model. Then n-best hypotheses obtained from this classifier were reranked using additional high-level features, such as number of coordinated adjective-noun and determiner-noun groups, number of correctly detected sentence clauses etc. They used logistic regression as reranking algorithms, which was trained on GICR data in order to assign higher score to correct sentence parses. To increase the quality of basic classifier tags in the training corpora were enriched with transitivity information for verbs and case label for nouns.

[5] https://drive.google.com/drive/folders/0B600DBw1ZmZASDFRVkJVd0pqNXM

IQMEN teams collected a set of hypotheses for each word and learnt the best one using the features for the word under consideration as well as for the word in a window of width 7 around it. Features included morphological (e.g part-of-speech, number, case, gender etc.) and graphical (suffixes, capitalization) information both for the word itself and its neighbours. The optimal tags for the sentence were guessed from left to right in a greedy fashion, instead of the tags for the words to the right the ambigiity classes were used. Similar approach was applied by Morphobabushka team, however. they refused to use any dictionaries guessing the tags for unknown words basing on their suffixes and features of surrounding words. IQMEN applied SVM with hash kernel, while Morphobabushka implemented SVM-NB classifier.

Sag team uses convolutional neural network, taking character-level representations for individual words and using several additional layers to comprise them into the representation of the whole sentence. Their algorithm does not use any dictionary except collected from the training set. Aspect team applies similar approach but use their own dictionary together with error-processing model to deal with typos and colloquial writing.

Table 3. Results of MorphoRuEval-2017

Team name	team ID	Track	Number of the best try	Accuracy by tags	Accuracyby sentences	Lemmatization, accuracy by wordforms	Lemmatization, accuracy by sentences
MSU-1	C	Closed	2	93.39	65.29		
IQMEN	O	Closed	1	93.08	62.71	92.22	58.21
Sagteam	H	Closed	2	92.64	58.40	80.73	25.01
Aspect	A	Closed	2	92.57	61.01	91.81	56.49
Morphobabushka	M	Closed	2	90.07	48.10		
Pullenti Pos Tagger	G	Closed	4	89.96	47.23	89.32	45.18
	B	Closed	6	89.91	48.2		
	N	Closed	4	89.86	47.13	85.10	29.04
	K	Closed	4	89.46	48.54	88.47	44.78
	F	Closed	2	88.14	39.63	87.27	36.90
	I	Closed	2	86.05	34.62		
	L	Closed	2	71.48	6.48		
ABBYY	E	Open	3	97.11	83.68	96.91	82.13
Aspect	A	Open	4	92.38	60.90	87.66	41.12
	N	Open	5	90.88	51.77	85.91	32.57
	J	Open	1	83.51	29.69		
	D	Open	5	77.13	17.19		

Neural networks approach are the clear winner, however, several remarks should be made. ABBYY team uses an additional corpus with rich annotation to train their model, it is not clear, whether their advantage would be so clear without it. On the closed task neural network methods are slightly behind more linguistically oriented approaches based on linear classifiers with rich feature descriptions. Therefore it is reasonable to ask, if neural network approach has the same benefits when only limited amount of training data is available. Interestingly, that on the SIGMORPHON-2016

[Cotterell et al., 2016] competition on morphological reinflection the same pattern was observed: elaborated neural network approaches clearly outperformed more traditional ones which attempted to utilize more linguistically motivated features. Another reasonable question is whether algorithms of different type can be combined together to compensate their weaknesses, for example, MSU team method can take any classifier as the basic one provided it ables to generate n-best lists of hypotheses together with probability estimates.

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]). Note that training corpus included only LJ posts and test corpus contained texts from three different sources, so the top-performing systems also demonstrated its ability to perform successfully not only on the domain they were trained on, but also on the texts from different origin.

7. Problems and discussion

One of the purpose of the work was to provide a unified training corpus for morphological tagging containing texts from different sources. However, it was not realized in full. As already mentioned, different corpora have different standards of lemmatization, for example for pronouns (what is the lemma of *она* 'she', *он* 'he' or *она*), but more important is that they have different standards of morphological annotation. There are many border cases treated in a different way in different corpora, such as the distinction between adverbs, predicatives and short adjectives, processing of reflexive verbs (they belong to special medial voice in GICR and RNC while UD SynTagRus distinguishes only active and passive voices) and so on. All competition participants trained their model only on GICR subset of the training corpus, which demonstrates that even after conversion to the same format joint usage of corpora of different origin and genre structure is problematic. Therefore the problem of unification is far from its solution, however, UD format looks appealing to be a destination of conversion from other formats.

For Russian language there is no dictionary in UD format, which could be used by participants and organizers to verify their decisions. In the framework of MorphoRuEval, there was carried out some work by the organizing committee, on the development of the OpenCorpora.org open-source dictionary: the dictionary was expanded by several thousand paradigms from the GICR dictionary, and then converted to universal dependencies. We hope that this dictionary will be included in the official UD documentation for Russian and will be useful in future evaluation.

8. Conclusion

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;
- 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.
- dataset for future improvement of morphological parsers, comprising texts from different sources, was created.

All materials of MorphoRuEval-2017 including training and test set are now available at the competition's github[6]. We welcome NLP-researchers and specialists in machine learning to use this collection and we hope that the collection will stay practical and relevant for a long time.

Acknowledgements

Authors express sincere gratitude to all the competition participants and especially thank Mikhail Korobov and Anton Kazennikov for their help during format discussion and explicit determination of unification conditions. We thank ABBYY Compreno and GICR teams for automatic annotation of initial subset of GICR subcorpora and assistance in its conversion to competition format. We are also grateful to Ilia Karpov from Kvant, Svetlana Savchuk from RNC and members of OpenCorpora and GICR teams for providing and preparation of their resources.

The work was partially supported by the grant NSh-9091.2016.1 for leading scientific groups. The reported study is supported by RFBR grants № 15-07-09306 "Evaluation Standards in Automatic Information Retrieval for Russian" and № 16-29-09614 "Principles and technologies of machine processing for social network texts in Russian". The UD SynTagRus corpus preparation was supported by GA UK grant 794417.

References

1. *Brants T.* TnT: a statistical part-of-speech tagger. In Proceedings of the sixth conference on Applied natural language processing. — Association for Computational Linguistics, 2000. — Pp. 224–231.
2. *Cotterell Ryan, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden* (2016) The SIGMORPHON 2016 shared task—morphological reinflection. In Proc. of the 2016 Meeting of SIGMORPHON.
3. *Dereza O. V., Kayutenko D. A., Fenogenova A. S.* (2016) Automatic morphological analysis for Russian: A comparative study. In Proceedings of the International Conference Dialogue 2016. Computational linguistics and intellectual technologies. Student session (online publication). Retrieved from: http://www.dialog-21.ru/media/3473/dereza.pdf
4. *Huang Z., Xu W., Yu K.* Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991. — 2015. Retrieved from https://arxiv.org/abs/1508.01991

[6] https://github.com/dialogue-evaluation/morphoRuEval-2017

5. *Lyashevskaya, Olga, Irina Astaf'eva, Anastasia Bonch-Osmolovskaya, Anastasia Garejshina, Julia Grishina, Vadim D'jachkov, Maxim Ionov, Anna Koroleva, Maxim Kudrinsky, Anna Lityagina, Elena Luchina, Eugenia Sidorova, Svetlana Toldova, Svetlana Savchuk, and Sergej Koval'* (2010) NLP evaluation: Russian morphological parsers [Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka]. In: Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2010. Vol. 9 (16), 2010. Pp. 318–326.
6. *Müller T., Schmid H., Schütze H.* Efficient higher-order CRFs for morphological tagging. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. — 2013. — Pp. 322–332.
7. *Nivre J.* (2016) Reflections on Universal Dependencies. Uppsala University. Department of Linguistics and Philology.
8. *Nivre J., de Marneffe M.-C., Ginter F., Goldberg Y., Hajic J., Manning Ch. D., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D.* (2016), Universal Dependencies v1: A Multilingual Treebank Collection, Proc. of LREC 2016, Portorož, Slovenia, pp. 1659–1666.
9. *Schmid H.* Treetagger (1995) A language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,1995. — Vol. 43, pp. 28.
10. *Selegey D., Shavrina T., Selegey V., Sharoff S.* (2016) Automatic morphological tagging of russian social media corpora: training and testing.In: Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2016.
11. *Sha F., Pereira F.* Shallow parsing with conditional random fields. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology — Association for Computational Linguistics, 2003. — Vol. 1., pp. 134–141.
12. *Sichinava D. V.* Parts of speech. [Chasti rechi. Materialy dlja proekta korpusnogo opisanija russkoj grammatiki (http://rusgram.ru)]. Moscow, ms. 2011. Available at: http://rusgram.ru/Chasti_rechi.
13. *Straková J., Straka M., Hajic J.* Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In Proceedings of ACL (System Demonstrations). — Association for Computational Linguistics, 2014. — pp. 13–18.
14. *Toldova, S., Sokolova, Elena, Astafiyeva, Irina, Gareyshina, Anastasia, Koroleva, Anna, Privoznov, Dmitry, Sidorova, Evgenia, Tupikina, Ludmila, Lyashevskaya, Olga.* Ocenka metodov avtomaticheskogo analiza teksta 2011–2012: Sintaksicheskie parsery russkogo jazyka [NLP evaluation 2011–2012: Russian syntactic parsers]. In Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue 2012. Vol. 11 (18), 2012. Moscow: RGGU, pp. 797–809.
15. *Toutanova K.. Klein D., Manning C., Singer Y.* Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. — Association for Computational Linguistics, 2003. — Pp. 173–180.

Appendix

News

team ID	track	Accuracy by tags	Accuracy by sentences	Lemmatization, accuracy by wordforms	Lemmatization, accuracy by sentences
C	closed	93.71	64.80		
O	closed	93.99	63.13	92.96	56.42
H	closed	93.35	55.03	81.60	17.04
A	closed	93.83	61.45	93.01	54.19
M	closed	90.52	44.41		
G	closed	89.73	39.66	89.04	37.71
B	closed	90.79	43.58		
N	closed	91.53	49.16	87.01	25.70
K	closed	90.36	45.53	89.23	40.22
F	closed	90.43	36.87	89.61	33.52
I	closed	88.66	29.89		
L	closed	75.88	2.790		
E	open	97.37	87.71	97.18	85.75
A	open	93.83	61.45	88.35	33.24
N	open	91.98	52.51	87.20	27.93
J	open	84.25	23.18		
D	open	79.52	10.89		

VKontakte

team ID	track	Accuracy by tags	Accuracy by sentences	Lemmatization, accuracy by wordforms	Lemmatization, accuracy by sentences
C	closed	92.29	65.85		
O	closed	92.39	64.08	91.69	61.09
H	closed	92.42	63.56	82.80	35.92
A	closed	91.49	61.44	90.97	60.21
M	closed	89.55	51.41		
G	closed	89.17	54.58	88.65	52.64
B	closed	88.96	52.29		
N	closed	88.44	48.59	83.67	34.51
K	closed	88.39	52.11	87.34	48.94
F	closed	86.72	44.72	85.81	41.90
I	closed	84.29	41.73		
L	closed	70.13	14.61		
E	open	96.52	81.34	96.26	79.93
A	open	90.92	61.09	86.97	48.24
N	open	89.63	52.29	84.58	36.80
J	open	82.87	36.44		
D	open	75.42	23.42		

Modern literature

team ID	track	Accuracy by tags	Accuracy by sentences	Lemmatization, accuracy by wordforms	Lemmatization, accuracy by sentences
C	closed	94.16	65.23		
O	closed	92.87	60.91	92.01	57.11
H	closed	92.16	56.60	77.78	22.08
A	closed	92.40	60.15	91.46	55.08
M	closed	90.13	48.48		
G	closed	90.97	47.46	90.28	45.18
B	closed	89.98	48.73		
N	closed	89.61	43.65	84.61	26.9
K	closed	89.63	47.97	88.84	45.18
F	closed	87.26	37.31	86.39	35.28
I	closed	85.21	32.23		
L	closed	68.43	2.03		
E	open	97.45	81.98	97.3	80.71
A	open	92.40	60.15	87.65	41.88
N	open	91.02	50.51	85.95	32.99
J	open	83.42	29.44		
D	open	76.45	17.26		

Algorithm description

Team	Track	Achievements	Method	additional training set (for open track only!)	Dictionary
Pullenti	closed	3rd place by mean lemmatization accuracy (by wordforms and sentences) and on VK, Modern literature	Rule-based approach, no training set used	—	Own dictionaries
Mental Computing	closed	3rd place by lemmatization (wordforms) on News	Char-level neural networks using Keras. The core algorithm is a grid classifier built using a RNN on LSTM, training on GICR data.	—	Dictionary collected from the training set
Abbyy	open	1st place by all metrics on open track	Bidirectional LSTM with probabilities and features from Abbyy NLC module, converted to UD	Pre-training on a large corpus (several tens of millions of words, including Russian Wikipedia) tagged by Compreno, then learning on GICR training data with more accurate tagging and a more suitable genre components.	Abbyy Compreno Dictionary

Team	Track	Achievements	Method	additional training set (for open track only!)	Dictionary
Sag	closed	3rd place by mean accuracy (tags and sentences) and on VK	2-layer deep learning neural network A two-level representation of a sentence by individual characters level (see Section 2.1.1) and level of words (Section 2.1.2), inspired by works [Nogueira dos Santos C., Zadrozny B.], [Zhiheng H., Wei X., Kai Y.], [Plank B., Søgaard A., Goldberg Y.]. Keras framework	—	Dictionary collected from the training set
Aspect	closed	1st place by lemmatization accuracy on news,2st place by tag accuracy on news, 3rd place by mean lemmatization accuracy (by wordforms and sentences) on all segments, closed track	Deep neural networks (based on recurrent neural networks) with the char-level representation of words	—	Own dictionaries for spell-checking and internet-slang
Aspect	open	2nd place by all metrics on open track	Deep neural networks (based on recurrent neural networks) with the char-level representation of words	Own tagged corpora of internet-texts	Own dictionaries for spell-checking and internet-slang
KZN	closed	2nd place by by mean accuracy (by wordforms and sentences) on all segments, 1st place by mean accuracy on News, 1st place by mean lemmatization accuracy on all segments	The model consists of four parts: the morphology module based on the AOT dictionary and GICR corpus, the predictive morphology module on the basis of the corpus, the SVM-classifier for removing morphological homonymy, and the context-dependent procedure for tagging the whole sentence.	—	AOT dictionary
Biser	closed	3rd place by lemmatization accuracy (by sentences) and on News, Modern literature	dictionary-based morphological guesser, homonymy is resolved using CRF.	—	
MSU-1	closed	1st place by mean accuracy (by wordforms and sentences) on all segments, closed track	Baseline HMM model. features reflecting grammatical correctness for reordering , reordering is performed using logistic regression	—	Abbyy Compreno Dictionary