Morphological Analysis for Russian: Integration and Comparison of Taggers

Alexandra Blazhievskaya, Elizaveta Kuzmenko, Elmira Mustakimova, Timofey Arkhangelskiy, Svetlana Toldova

National Research University Higher School of Economics

Moscow, Russia

{aablazhievskaya\_2, eakuzmenko\_2, egmustakimova\_2}@hse.edu.ru, {tarkhangelskiy, stoldova}@hse.ru

*Abstract* — In this paper we present a comparison of three morphological taggers for Russian with regard to the quality of morphological disambiguation performed by these taggers. We test the quality of the analysis in three different ways: lemmatization, POS-tagging and assigning full morphological tags. We analyze the mistakes made by the taggers, outline their strengths and weaknesses, and present a possible way to improve the quality of morphological analysis for Russian.

# Introduction

In this paper we present the results of testing different morphological taggers for the Russian language. Russian is a highly inflective and morphologically rich language, and developing decent morphological tools for Russian presents a huge problem even for advanced researchers.

A considerable number of taggers provide morphological disambiguation while performing POS-tagging for Russian, however all of them are erroneous in some way. This disadvantage can be beneficial since the taggers make errors in different issues: when one analyzer fails, another may guess the correct tag. Therefore, it could be very useful to inspect the performance of each tagger and reveal the specificity of the mistakes it makes. These findings can then help to build an improved tagger for Russian that will combine in itself all the forces of other taggers. The near future of morphological analysis of Russian, as we see it, is meta-learning, in which all the cases where taggers guess tags correctly are taken and all the cases where the taggers make errors are omitted.

The question is then: do the cases where taggers make errors overlap or not? We answer this question in our paper via the experiment in which we build a gold standard corpus and compare the tags found in this corpus to those that are output by our taggers. In case of discrepancy, we analyze the cause of an error.

The structure of this paper is as follows. In Section II we describe the previous work in this field: how the standards for morphology annotation were defined and what the specific morphology problems for the Russian language are. We also give an overview of the instruments developed for Russian: taggers *Freeling*, *Pymorphy*, *MyStem* and *TreeTagger*, and describe previous attempts to compare their performance. Section III presents an experiment of comparing the taggers: we analyze the differences in the tagsets and define the rules for the unification of morphological tags. Section IV provides the results of our experiment, and in Section V we discuss these results and propose the way towards organizing meta-learning of the taggers.

II. Background

*A. Standards for morphology annotation for Russian. Morphology problems.*

The Russian language presents certain problems with regard to morphology annotation, because it is a highly inflectional language with many grammatical categories. There is no standard even for part-of-speech annotation, let alone subtle categories such as (im)perfectiveness and animacy. Theoretical disputes concerning Russian morphology lead to variety of solutions for morphology annotation – from positional tags following the MULTEXT-East guidelines [1] to combinations of tags employed in RNC [8]. An additional problem arises from the fact that tags in Russian are various and therefore can be combined and simplified in different ways (some systems do not account for one or another grammatical category)

*B. Disambiguation for Russian. Gold Standard sets.*

Until recently, there were no standard golden corpus for Russian, and even no unification for morphology annotation. Presently, there are two corpora that could serve as models for annotation tasks: a disambiguated RNC subcorpus and Opencorpora [2]. In addition, a RU-EVAL shared task guideline proposes unification rules for the output of different morphological taggers and a gold standard corpus consisting of 3 thousand word tokens.

The difficulties in the different taggers output unification come from the high variability of tagsets. The comparison of taggers for Russian is also complicated by the fact that there are different traditions for the lemmatization process. For example, some taggers count for verbs as lemmas of participles, and other taggers lemmatize participles as adjectives.

*C. Taggers for Russian*

*1) MyStem* [3] is a morphological analyzer with disambiguation developed for the Russian language by Ilya Segalovich and Vitaliy Titov at “Yandex”. In the core of the software lies a dictionary that helps generate morphological hypotheses for both known and unknown words.

*2) Pymorphy2* [4] is a morphological analyzer developed for the Russian language by Mikhail Korobov on the basis of OpenCorpora dictionaries. PyMorphy2 is written fully in the Python programming language and is able to normalize, decline and conjugate words, provide analyses or give predictions for unknown words.

*3) Freeling* [5] is a set of open source linguistic analyzers for several languages. It features tokenizing, sentence splitting, morphology analyzers with disambiguation, syntax parsing, named entity recognition, etc. In this research, we use only morphological analyzer for Russian.

*4) TreeTagger* [6, 7] is a language independent part-of-speech tagger developed by Helmut Schmid. TreeTagger is based on decision trees and should be trained on a lexicon and a manually tagged training corpus. The program can annotate texts with part-of-speech and lemma information.

III. Experiment design

In the experiment we evaluated the performance of our taggers compared to the gold standard corpus.

As a gold standard we took a part of a disambiguated subcorpus of Russian National Corpus [8]. This subcorpus contains 5.9 million tokens, annotated morphologically with the help of *MyStem* and disambiguated by hand. Therefore, the tokens have only one morphological analysis, and the tagset in this corpus complies to the one developed for *MyStem*. We took randomly a 1/20 part of this corpus, containing 426 thousand tokens to serve as a training set.

One of the problems in our experiment is that all analyzers have different notations for parts of speech and morphological categories. The discrepancies between the tagsets can be of different kinds:

* Some morphological category is present in the tagset of the gold standard but absent in the tagset of another morphological analyzer: for example, *MyStem* distinguishes between animate and inanimate nouns or adjectives as it has specific dictionaries where these characteristics are defined for every word. *TreeTagger*, however, does not consider this feature to be important for morphology and does not include it in the analysis.
* Morphological analyzers have different standards concerning part of speech identification: for example, *Freeling* identifies participles as a separate part of speech, whereas other morphological analyzers identify participles as verbal forms.
* Consequently, alongside with different standards towards part of speech identification, parsers assign different lemmas to tokens problematic in this aspect: therefore, the lemma for the word ‘*сделанной*’ (participle) would be ‘*сделанный*’ (participle) in *Freeling* and ‘*сделать*’ (verb) in  *MyStem*.
* If all the taggers provide the same part of speech for a particular word, there still can be problems with lemmatization: for example, *TreeTagger* assigns one and the same aspect to Russian verbs in different aspects, and so does Freeling, but other tagsets do not require the aspect to be changed in the process of lemmatization.

Due to these problems, we need to define conventions that will enable us to compare the taggers’ output. Since our gold standard is annotated by *MyStem*, we decided to convert the tags of all analyzers into *MyStem*–styled tags. The rules of conversion are presented in Table 1.

TABLE I. Rules for conversion of all tags into *MyStem*-styled tagset

|  |  |
| --- | --- |
| **Gold standard tag** | **Tag counted as correct** |
| A-NUM (numeric adj.) | NUM (numeral) |
| PARENTH(parenthesis) | ADV(adverb) |
| ADV-PRO (adv.-pronoun) | PRO (pronoun) |
| A-PRO (adj.-pronoun) | PRO (pronoun) |
| m-f (common gender) | both are correct |
| anim (animacy) | not important |
| inan (inanimateness) | not important |
| dat2 (the 2nd dative) | dat (dative) |
| gen2 (the 2nd genitive) | gen (genitive) |
| acc2 (the 2nd accusative) | acc (accusative) |
| loc2 (the 2nd locative) | loc (locative) |
| adnum (count form) | NUM (numeral) |
| intr (intransitiveness) | not important |
| tran (transitiveness) |  not important  |

However, Table 1 contains rules only for the least problematic cases. More problematic cases include, as it was mentioned earlier, lemmatization of participles and perfective verbs. These issues we solve by assigning lemmas given by *Mystem* and taking the tag itself from another analyzer. In addition, we do not consider identifying patronyms, zoonyms and other lexical classes to be of importance for the task of morphological analysis and exclude them from our experiment.

IV. Evaluation

Before evaluation the gold standard text was analyzed by the three taggers. Their outputs were converted into a single tagset as described in the previous section. After such preprocessing, we could start evaluating the analyses.

For each word we compared the analyses of the three taggers and the analysis presented in the gold standard. In particular, we checked whether the part of speech was the same and if the set of grammatical categories contained in the tag was identical to the gold standard. There were three major modes of evaluation:

1) checking the correspondence between the assigned lemmas;

2) checking the correspondence between the assigned parts of speech;

3) checking the correspondence between the assigned morphological tags in the whole.

Table 2 presents the results for all our taggers and three modes.

TABLE II. Results of the comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tagger** | **Mode** | **Precision** | **Recall** | **F1** |
| Freeling | Lemma | 81.98% | 100% | 0.902 |
| POS | 91.19% | 98.76% | 0.948 |
| Full tag | 81.31% | 100% | 0.897 |
| Pymorphy2 | Lemma | 87.8% | 100% | 0,95 |
| POS | 90.5% | 99.3% | 0.947 |
| Full tag | 59.2% | 100% | 0.743 |
| TreeTagger | Lemma | 97% | 93% | 0.950 |
| POS | 95% | 97% | 0.960 |
| Full tag | 91% | 98% | 0.944 |

As it is shown in Table 2, all the taggers present decent results, but none of them performs without mistakes. The best results are received for the *TreeTagger*. *Pymorphy2* also presents high precision and recall for all modes except complete tags.

V. Conclusion

In this paper we presented an analysis of the performance of three taggers for Russian. The received results are of interest to anyone engaged in morphological analysis of Russian. As a future step we plan to build a meta-learning system based on several taggers. Such system will take as input the morphological analyses from several taggers, identify which tagger provides the best guess for each particular case, and give as output the combination of correct variants. We expect this system to be highly accurate.

References

1. Erjavec, T. (2004, May). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *LREC*.
2. Bocharov, V., Bichineva, S., Granovsky, D., Ostapuk, N., & Stepanova, M. (2011). Quality assurance tools in the OpenCorpora project. In *Computational Linguistics and Intelligent Technology: Proceeding of the International Conference «Dialog* (pp. 10-17).
3. MyStem, Web: https://tech.yandex.ru/mystem/.
4. PyMorphy2, Web: https://pymorphy2.readthedocs.org/en/latest/.
5. Padró, L., & Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality.
6. H. Schmid, “Improvements in Part-of-Speech Tagging with an Application to German”, *Proceedings of the ACL SIGDAT-Workshop*. 1995.
7. H. Schmid, “Probabilistic Part-of-Speech tagging using decision trees”, *Proceedings of International Conference on New Methods in Language Processing*. 1994.
8. Russian National Corpus, Web: http://ruscorpora.ru/en/.