RusPoS. Part-of-Speech Tagger for Russian language

Author
Rinat Gumarov
Innopolis University

Supervisor
Leon Derczynski
Innopolis University

Abstract

In this paper we propose a PoS Tagger for Russian. There are about 260 million native speakers. We find russian corpus which is consists of 1,966,619 tagged words. We develop features, train classifier and report tagging results is 90% accuracy on classifier, trained on 1% of corpus. The data and tools are in open-source on github

1 Introduction

There are a lot of information, written on Russian language on the Internet. And we need tools to understand and exploit these data. One of the most fundamental parts of the linguistic analysis is part-of-speech tagging, a basic form of syntactic analysis which has countless applications in NLP.

In this paper, we produce a Russian POS tagger. Our contributions are as follows:

- we find and parse a POS tagged texts on Russian language,
- we developed features for Twitter POS tagging and conducted experiments to evaluate them
- we provide our trained POS tagger to the research community.

Tag	Example	Description
PNCT	«	Any punctuation
NOUN	Школа	nouns
VERB	учит	verbs
INFN	прикусить	verbs, infinitive form
PRCL	ли	particle
PREP	в	preposition
ADJF	новом	full adjective
NPRO	это	pronoun-noun
ADVB	уже	adverb
UNKN	ребрендинг	unknown word(not in dictionary)
PRED	можно	predicate
CONJ	что	conjunction
COMP	дальше	comparative
PRTF	появившихся	full participle
NUMR	два	numeral
INTJ	Однако	interjection
PRTS	посвящен	short participle
GRND	будучи	gerund
ADJS	долго	short adjective
ROMN	XVII	Roman numeral
NUMB	12	numbers
LATN	deus	Latin words
SYMB	+	symbols

This was made possible by a feature set that captures Russian language specific properties. The success of this approach demonstrates that with careful design, supervised machine learning can be applied to rapidly produce effective language technology in new domains.

2 Russian language

Russian language is a language with strong rules for most of parts of speech. The main rule of word formation is so called 'окончания'(endings) or in other words - suffixes. For example verbs often end with suffixes, like: 'ться', 'ся', 'л', 'ть', 'ешь', 'ет', 'ем', 'ете', 'ут' '-ют', 'ишь', 'ит', 'им', 'ите', 'ат' 'ят'. The main problem is that there are also not

verbs, which could end with such suffixes. Here comes order of the words in the sentence. Order of parts of speech is also important rule, but not as important as in English. There are also could be identically written sentences, which have different meaning. For example, 'Эти типы стали есть на складе' which could be translated either as 'These types of steel are in stock' or 'These people began to eat in the warehouse'. In first case this sentence would be tagged as 'Эти/NPRO типы/NOUN стали/NOUN есть/VERB на/PREP складе/NOUN', and in second case: 'Эти/NPRO типы/NOUN стали/VERB есть/VERB на/PREP складе/NOUN'. In this case problem could be solved by defining context. The other issue is that one word could have different part of speech in one sentence: 'Косил косой косой косой' ('an oblique man mowed the grass with spit'); tagget : 'Косил/VERB косой/NOUN косой/ADJ косой/NOUN'. This problem is unsolvable, because in the Russian language the order of parts of speech may be different.

3 System

As a classifier we used DecisionTreeClassifier. The goal of the decision tree is to create a model that predicts the meaning of the goal.
Benefits:
- Only a little data preparation is required. Other methods often require data to be normalized, you need to create dummy variables and delete empty values.
- Ability to process both numeric and categorical data.

Disadvantages:
- Decision tree students can create complex trees that do not generalize data.
- Trees of decision-making can be unstable due to small differences in the data.

Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the decision tree.

To train the model next set of features were used:
- word itself
- is first word in sentence
- is last word in sentence
- is capitalized
- 1 char prefix
- 2 chars prefix
- 1 char suffix
- 2 chars suffix
- 3 chars suffix
- 4 chars suffix
- end with vowel
- vowels in word
- is token in punctuation signs
- previous word
- is first char is a latin letter
- previous word 1 char suffix
- previous word 2 chars suffix
- previous word 3 chars suffix
- next word 1 char suffix
- next word 2 chars suffix
- next word 3 chars suffix
- is token contains hypen
- is token has capitals inside

4 Experiments

Our evaluation was designed to test the efficacy of this feature set for part-of-speech tagging given limited training data. We divided the set of 1210 annotated sentences into a training set of 907 (17,793 tokens), and a test set of 303 (8,292 tokens). The results are shown in table below.

tag	precision
ADJF	0.84
ADJS	0.70
ADVB	0.83
COMP	0.38
CONJ	0.97
GRND	0.60
INFN	0.93
INTJ	0.78
LATN	0.67
NOUN	0.88
NPRO	0.71
NUMB	0.92
NUMR	0.73
PNCT	0.99
PRCL	0.95
PRED	1.00
PREP	0.91
PRTF	0.64
PRTS	0.75
ROMN	1.00
SYMB	1.00
UNKN	0.44
VERB	0.95
avg / total	0.89

Based on the test results, it is possible to identify that our pos tagger works bad with comparative (that is because comparative is similar to adjective and adverb, also testing data is not huge enough) and unknown words(obviously, our pos tagger can not find these words, because it does not have russian dictionary, and even if word is not in dictionary it is very similar to known words even for people).

5 Conclusion

We have developed a part-of-speech tagger for Russian language and have made tools available to everyone at github. More generally, we believe that our approach can be applied to address other linguistic analysis needs.

6 References

[1] Russian language rules.
Available: https://best-language.ru

Sklearn. Machine learning kit.
Available: http://scikit-learn.org/stable/documentation.html