Communications in Computer and Information Science 2015 vol.505, pages 263-275

A comparative evaluation of statistical part-of-speech taggers for Russian

Gareev R., Ivanov V.

Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

© Springer International Publishing Switzerland 2015. Part-of-speech (POS) tagging is an essential step in many text processing applications. Quite a few works focus on solving this task for Russian; their results are not directly comparable due to the lack of shared datasets and tools. We propose a POS tagging evaluation framework for Russian that comprises existing third-party resources available for researchers. We applied the framework to compare several implementations of statistical classifiers: HunPos, Stanford POS tagger, OpenNLP implementation of MaxEnt Markov Model, and our own reimplementation of Tiered Conditional Random Fields. The best tagger that was trained on a corpus with less than one million words achieved an accuracy above 93% .We expect that the evaluation framework will facilitate future studies and improvements on POS tagging for Russian.

http://dx.doi.org/10.1007/978-3-319-25485-2_8