Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2013 vol.7816 LNCS NPART 1, pages 329-342

Introducing baselines for Russian named entity recognition

Gareev R., Tkachenko M., Solovyev V., Simanovsky A., Ivanov V.
Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

Current research efforts in Named Entity Recognition deal mostly with the English language. Even though the interest in multi-language Information Extraction is growing, there are only few works reporting results for the Russian language. This paper introduces quality baselines for the Russian NER task. We propose a corpus which was manually annotated with organization and person names. The main purpose of this corpus is to provide gold standard for evaluation. We implemented and evaluated two approaches to NER: knowledge-based and statistical. The first one comprises several components: dictionary matching, pattern matching and rule-based search of lexical representations of entity names within a document. We assembled a set of linguistic resources and evaluated their impact on performance. For the data-driven approach we utilized our implementation of a linear-chain CRF which uses a rich set of features. The performance of both systems is promising (62.17% and 75.05% F1 measure), although they do not employ morphological or syntactical analysis. © 2013 Springer-Verlag.

http://dx.doi.org/10.1007/978-3-642-37247-6_27