International Journal of Applied Engineering Research 2015 vol.10 N24, pages 44795-44800

Methods and software tools of morphological disambiguation in the texts in tatar

Gataullin R., Gilmullin R., Suleymanov D.

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

Abstract

© Research India Publications. This article provides a review of analytical methods for resolving the problem of morphological ambiguity and analysis of their applicability to the Tatar language. Since the task was set still in the 50-60-ies of XX century, the methods of solution have been accumulated quite a lot. Basically they can be divided into methods of rule-based and statistical and probabilistic methods. Methods are mainly language independent, each has its advantages and disadvantages, and their accuracy varies from one language to another. For example, for the English language, which has a poor morphology and the fixed order of the words, the accuracy reaches 94-96%. And for the Russian language with free word order, such accuracy is difficult. To resolve the ambiguity in morphological Tatar language in terms of the characteristics of the language such as agglutinative feature and free word order, it is offered a fusion of these methods, by which a high precision resolution is supposed to be achieved. At the moment, the research is still in progress, the tools for the development of contextual rules have been designed, subcorpus for statistical machine learning and probabilistic models is also being elaborated. In addition to the methods, the article describes the current state of the electronic corpus of the Tatar language, and discusses the problems and possible solutions to the problem of polysemanticism in the corpus markings.

Keywords

Electronic corpus of the language, Resolving morphological ambiguity, The Tatar language