Towards the Uzbek Language Endings as a Language Resource

Sanatbek Matlatipov1(B) , Ualsher Tukeyev2 , and Mersaid Aripov1

1 National University of Uzbekistan, University Street, 14, 100174 Tashknet, Uzbekistan
mr.sanatbek@gmail.com, mirsaidaripov@mail.ru
2 Al-Farabi Kazakh National University, Al-Farabi Ave., 71, 050040 Almaty, Kazakhstan
ualsher.tukeyev@gmail.com

Abstract. The Uzbek language belongs to low-resource languages. It is very important to increase the number of language resources such as dictionaries, corpora (monolingual and bilingual) for the Uzbek language. Dictionaries may be different kinds: monolingual, orthographical, bilingual, grammar special dictionaries: stems dictionaries, affixes dictionaries, etc. For different NLP tasks of agglutinative languages, such as morphological analysis, information retrieval, machine translation (segmentation preprocessing) in some cases needs a dictionary of words' endings. In this paper, we proposed the first electronic dictionary of Uzbek words' endings in variants for morphological segmentation preprocessing useful for neural machine translation. The resource analysed by the initial version of the Lexicon free stemming tool [3] created by authors. For creation of Uzbek words' endings' electronic dictionary, it was used a combinatorial approach inferring apply for part of speech of the Uzbek language: nouns, adjectives, numerals, verbs, participles, moods, voices.

Keywords: Uzbek language · Electronic dictionary · Words' endings · Language resource · Lexicon free stemming tool