Inferring the complete set of Kazakh endings as a language resource

Ualsher Tukeyev [0000-0001-9878-981X] and Aidana Karibayeva [0000-0002-2023-1573]

Al-Farabi Kazakh National University, al-Farabi ave., 71, 050040, Almaty, Kazakhstan
ualsher.tukeyev@gmail.com, a.s.karibayeva@gmail.com

Abstract. The Kazakh language belongs to low-resource languages. For application of actual modern branches as artificial intelligence, machine translation, summarization, sentiment analysis, etc. to the Kazakh language needs increasing the number of electronic language resources. Although neural machine translation (NMT) has shown impressive results for many world languages, it does not solve the problem of low-resource languages. Therefore, the development of resources and tools perfecting the use of NMT for low-resource languages is relevant. For perfect use of NMT for the Kazakh language needs bilingual parallel corpora, but also needs a perfect method of the segmentation source text. By the opinion of authors, one of the effective ways for source text segmentation is morphological segmentation. The authors propose to use for morphological segmentation of Kazakh text a table of a complete set of Kazakh words' endings. In this paper is described the inferring of the complete set of Kazakh words' endings. Development of the table of the complete set of word' endings of the Kazakh language will allow in one step (by reference to the table of endings of the language) to perform the segmentation of the word's ending into suffixes. The complete set of endings of the Kazakh language allows guaranteeing the analysis of any word of the Kazakh language, as this is determined by the inferring of the complete system of words' endings of the language.

Keywords: the Kazakh language, morphological segmentation, words' endings, language resource.