Abstract

The Georgian language has a difficult verbal system. To help foreigners learn Georgian, a linked-data base of inflected forms of Georgian verbs is being built: KARTUVERBS. It currently contains approximately 5 million inflected forms related to more than 16 000 verbs. There are more than 80 million links in the base. Response times are acceptable even when this linked-data base runs on a private machine. Data are generated from external textual source. We present a process to extract, improve, validate and convert textual structured knowledge into semantic linked data applied to Georgian verbal knowledge. The process successively applies improvement tools. In particular, we trained a classification model with supervised machine learning algorithm, decision tree, for predicting occasional missing values of verbal nouns, which typically represent lemmas in Georgian dictionaries. The model shows a precision of 98%-99%. This result is backed up by a comparison with the lemmas of a national Georgian corpus. Both our system and the corpus consistently agree when the corpus provides a lemma. For other cases, we set up an academic crowdsourcing platform called HEADWORK to gather user-suggested data and expertise from professionals of the domain. Participants can vote on the correctness of verbal nouns in relation to the given forms and leave comments and suggestions. For Georgian learners, KARTUVERBS is a digital lexicographic tool that assists in identifying verbal nouns from any inflected form and offers an API for translation. It it useful for understanding the Georgian language, particularly verb conjugation. For Georgian lexicographers, KARTUVERBS is the system to digitally contain almost all Georgian verbs and their inflected forms, serving both pedagogical and scientific purposes due to the complex and exception-laden nature of Georgian verb conjugation. For lexicographers of other languages, KARTUVERBS exemplifies a tool built on Semantic Web standards and technologies, making its data machine-readable and interoperable on the internet. The programs/scripts produced so far, for data extraction, improvement and validation are freely available.

Key words: Linked-data, machine learning, decision tree, Georgian verbs, dictionary lemmas.