Comparison between Humans and Machines on the Task of Accentuation of Slovene Words

Tomaž Šef, Tea Tušar, Andrej Bratko, Matjaž Gams
Department of Intelligent Systems
Jozef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia

The accentuation of unknown Slovene words represents a challenging task for automated solvers since in Slovenian, stress can be located on arbitrary syllables. Most words have only one stressed syllable, but there exist also words with no stress and words with more than one stress. Furthermore, different forms of the same word can be stressed differently. In this work, we inspect the performance of humans (human volunteers, human experts and expert-defined rules) and machines (machine learning methods and n-gram Markov models) on the task of stress assignment of Slovene words. We analyze the relation between human knowledge and their rules and compare the results of expert-defined rules and machines to find the best method for automatic accentuation of Slovene words. We find that humans tend to accentuate the words correctly, even when they have never heard or seen them before. On the other hand, expert-defined rules for accentuation perform quite poorly, achieving worse results than machines. This indicates that humans are good at accentuating, but very limited when their knowledge has to be formalized. Therefore, machine methods have to be employed for automatic accentuation of Slovene words.

Naglaševanje nepoznanih slovenskih besed: primerjava med človekom, človeškimi pravili in strojnim učenjem

Tomaž Šef, Tea Tušar, Andrej Bratko, Matjaž Gams
Odsek za inteligentne sisteme
Institut "Jožef Stefan"
Jamova 39, 1000 Ljubljana, Slovenija

Avtomatsko naglaševanje nepoznanih slovenskih besed je eden od zahtevnejših problemov pri razvoju različnih govornih sistemov. Za razliko od nekaterih drugih jezikov je za slovenski jezik značilno prosto mesto naglasa. Poleg tega ima lahko posamezna beseda različno število naglasnih mest. Mesto naglasa je določeno za vsako besedo posebej in velja, da se ga naučimo hkrati z učenjem jezika. V tem delu prikazujemo rezultate raziskave o sposobnostih ljudi (prostovoljci, jezikoslovci, človeška pravila) in strojev (metode strojnega učenja, n-gramski Markovski modeli) glede naglaševanja njim nepoznanih besed. Analiziramo povezavo med človeškim znanjem in pravili ter rezultate človeških pravil primerjamo s strojno generiranimi pravili. Na takšen način skušamo priti do najboljše metode za avtomatsko naglaševanje nepoznanih slovenskih besed. Ugotavljamo, da ljudje besede izgovarjajo večinoma pravilno, čeprav jih pred tem niso še nikoli slišali ali prebrali. Po drugi strani pa človeška pravila za naglaševanje delujejo nezanesljivo in dajejo slabše rezultate od računalniško generiranih pravil. Ljudje se razmeroma dobro obnesejo pri naglaševanju besed, ko pa je potrebno to znanje formalizirati, so rezultati nezadovoljivi. Zato je za avtomatsko naglaševanje besed bolje uporabiti metode strojnega učenja.