საკუთარი სახელების იდენტიფიცირების პრობლემა ქართულენოვანი ტექსტების დამუშავებაში

მ.ხაჩიძე, მ.ცინცაძე, მარჩუაძე, გ.ბესიაშვილი

ელ-ფოსტა: {manana.khachidze, magda.tsintsadze, maia.archuadze, gela.besiashvili}@tsu.ge,

კომპიუტერულ მეცნიერებათა დეპარტამენტი,
ზუსტ და საბუნებისმეტყველო მეცნიერებათა
ფაკულტეტი, ივ.ჯავახიშვილის სახელობის
თბილისის სახელმწიფო უნივერსიტეტი
თბილისი. ჭავჭავაძის გამზ. 3

ბუნებრივი ენის დამუშავების ამოცანების უმრავლესობა მოიცავს ტექსტების საწყისი დამუშავების პროცესს, რომელიც სხვადასხვა პროცედურებთან ერთად („სტოპ“ სიტყვების ამოყრა, სტემინგი, ლემატიზაცია) მოიცავს საკუთარი სახელების იდენტიფიცირებას. საკუთარი სახელების იდენტიფიცირება ასევე მნიშვნელოვანია დიდი მონაცემების ერთ-ერთი მთავარი წყაროს - სოციალური ქსელების ინფორმაციის დამუშავებისას. ეს ამოცანა შედარებით ადვილი გადასაჭრელია ისეთი ენებისათვის, რომელთა დამწერლობა და მართლწერის წესები უზრუნველყოფენ საკუთარი სახელების გარკვეული წესით (სახელი იწყება „დიდი“ ასოთი) ფორმალიზებას, მაგრამ ეს ამოცანა პრობლემურია ქართული ენის და მისი მსგავსი სხვა აგლუტინაციური ენებისათვის.

განვიხილავთ ქართული ტექსტებიდან საკუთარი სახელების ამოღების ალგორითმს. ალგორითმი წარმოადგენს ცნობილი „ბულის ძებნის“ მოდიფიცირებულ ვარიანტს, და იყენებს საკუთარ სახელების ბაზას. ალგორითმის წარმატებულობა დამოკიდებულია ბაზის სისრულეზე. გარდა ამისა, პროცედურის სრულყოფილი ციკლი მოიცავს ქართული ენის სტემინგის ალგორითმის გამოყენებას. ფუძის ამოღების ეს ალგორითმი სპეციფიურია და მორგებულია ქართულ ენაზე. ეს ფაქტი თავის მხრივ განსაზღვრავს ტექსტიდან საკუთარი სახელების ამოღების ალგორითმის ორიგინალურობას და ეფექტურობას ქართული ენისათვის.

ლიტერატურა

[1] Riyad Al-Shalabi, Ghassan Kanaan, Bashar Al-Sarayreh, Khalid Khanfar, Ali Al-Ghonmein, Hamed Talhouni, and Salem Al-Azazmeh. Proper Noun Extracting Algorithm for Arabic Language. International Journal of the Computer, the Internet and Management Vol. 19. No.1 (January-April, 2011) pp 45 -53

[2] Wesley W. Chu. Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities. Springer Science & Business Media, Sep 24, 2013 - Computers - 311 pages

[3] Ali Reza Ebadat, Vincent Claveau, Pascale Sebillot. Proper Noun Semantic Clustering Using Bag-of-Vectors. Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference. Marco. May 23, 2012 – May 25, 2012. Pp. 220-225.

.