ქართული ენის GeWordNet ლექსიკონისთვის ჰიპონიმური ხის ავტომატური ფორმირების ალგორითმი

ლიანა ლორთქიფანიძე

ელ-ფოსტა: liana.lortkipanidze@tsu.ge

პრაქტიკული ინფორმატიკა, კომპიუტერული მეცნიერების დეპარტამენტი, ზუსტი და საბუნებისმეტყველო ფაკულტეტი, ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი, თბილისი, უნივერსიტეტის ქ.13

ინტერნეტ სივრცეში საძიებო სისტემების ინტელექტუალიზაცია მნიშვნელოვნად ზრდის ძიების სიჩქარესა და ხარისხს. დოკუმენტებში ძიებისას ბუნებრივ წინააღმდეგობას ქმნის სინონიმია (სხვადასხვა ცნების აღნიშვნა ერთი და იგივე სიტყვით ან ტერმინით) და პოლისემია (საერთო სემანტიკის მქონე ცნებების აღნიშვნა სხვადასხვა სიტყვით ან ტერმინით). ბოლო წლებში ეს პრობლემა ბევრი ენისთვის დაძლეულია სპეციალური ელექტრონული WordNet ტიპის თესაურუსების გამოყენებით [1].

რუსთაველის ეროვნული ფონდის მიერ დაფინანსებული პროექტის ფარგლებში იქმნება ქართულ სიტყვათა ქსელის კომპაილერი - GeWordNet, რომლის ანალოგი საქართველოში ჯერჯერობით არ არსებობს. GeWordNet თესაურუსის გამოყენება შესაძლებელი გახდება:

- ინფორმაციის ძიებისას მომხმარებლის მოთხოვნის გასაფართოებლად პარადიგმატულად და სინტაგმატურად დაკავშირებული სიტყვების მეშვეობით. ასეთი სიტყვებია, მაგალითად, სინსეტის (SynSet [1]) კომპონენტები, ან „ზმნა-აქტანტი“-ს ტიპის კავშირები, რომლებიც კონტექსტური ძიებისათვის არის საჭირო;
- ფორმალური გრამატიკების ლექსიკონად, განსაკუთრებით ზმნების ვალენტობის, არსებითი და ზედსართავი სახელების ამომწურავი აღწერისას;
- სპეციალიზებული ლექსიკონების (მაგალითად, სამედიცინო, ეკონომიკური, გეოგრაფიული, ბიოლოგიური და სხვ.) შესადგენად;
- სხვადასხვა დიალექტებისა და ენების ლექსიკონების შესადგენად;
- სიტყვათა სინტაგმატური მიმართებების საშუალებით კლასიკური ამოცანის - სიტყვების არაერთმნიშვნელოვნების მოსახსნელად;
- ტექსტის ავტომატური დამუშავებისა და ინფორმაციული ძიების პროგრამულ დანართებში დოკუმენტების ფილტრაციისა და რუბრიკაციის ხარისხის გასაზრდელად;
- ჰიპერონიმული მიმართებების საფუძველზე აზრობრივად ახლო მდგომი ტექსტების განსაზღვრისთვის.

WordNet-ის ქართული ვერსიის რეალიზაციისათვის ამჟამად წარმოებს ქართული ენის არსებული ლექსიკონებიდან GeWordNet თესაურუსის ავტომატური კომპილირების პროცედურები [2]. მუშავდება ალგორითმი ლექსიკონის ჰიპონიმური ხის ასაგებად

[1] SynSet - სინონიმური მწკრივი, რომელშიც გაერთიანებულია მსგავსი მნიშვნელობის მქონე სიტყვები.

თითოეული სინონიმური მწკრივისთვის შერჩეული პროტოტიპისა და თესაურუსის სტრუქტურის მიხედვით. თესაურუსის ცნებათა შორის ჰიპონიმური და მერონიმური კავშირების დამყარება მთავარ ამოცანას წარმოადგენს ენათა შორის შესატყვისობის ინდექსის (ILI) აგების დროს [3]. მოხსენებაში განხილული იქნება ILI ინდექსის აგების დროს ჰიპონიმური ხის ავტომატური ფორმირების ალგორითმი.

ლიტერატურა

[1] Fellbaum C. (ed.), WordNet: An Electronic Lexical Database, MIT Press, 1998.

[2] Karlgren J. and Sahlgren M., From Words to Understanding // In Uesaka Y., Kanerva P., Asoh H. (eds.) Foundations of Real-World Intelligence. – Stanford: CSLI Publications, 2001. – P. 294 – 308.

[3] Salton G., Automatic Text Processing: The Transformation, Analisys, and Retrieval of Information by Computer. – Addison-Wesley, Reading, MA. – 1989.