არსებული ბარიერები ქართული ტექსტების ავტომატური დამუშავებისთვის

რუდოლფ ერემიან
თბილისის სახელმწიფო უნივერსიტეტის ზუსტ და საბუნებისმეტყველო ფაკულტეტის კომპიუტერულ მეცნიერებათა მიმართულების ბაკალავრიატის VII სემესტრის სტუდენტი
სამეცნიერო ხელმძღვანელი: ასისტ. პროფ. ლიანა ლორთქიფანიძე

NLP არის ხელოვნური ინტელექტის ერთერთი მიმართულება, რომლის მთავარი მიზანია ნატურალური ენის დამუშავება და გაგება, ჩვენს შემთხვევაში ქართული ენის. დღეს, ძალიან დიდი რაოდენობა ისეთი პოპულარული აპლიკაციების, როგორებიცაა, მაგალითად Google Search, Apple Siri, Amazon Stroe, იყენებენ NLP მეთოდებს ინფორმაციის ძებნისთვის - Information Retrieval, ხმის ამოცნობისათვის - Voice Recognition, ხმის სინთეზისათვის - Text-to-Speech, დიდ ტექსტებში საკვანძო სიტყვების ძებნისათვის - Topic Modelling და ა. შ.

არსებობს დიდი რაოდენობა სხვადასხვა ინსტრუმენტებისა, რომლებიც გვაძლევენ ტექსტების ავტომატური დამუშავების საშუალებას, მაგალითად NLTK Tool Python ენაზე ტექსტის ანალიზის შემთხვევაში და StanfordNLP თუ ვიყენებთ Java-ს. ამ ინსტრუმენტებს აქვთ ისეთი ბუნებრივი ენების მხარდაჭერა, როგორიცაა: ინგლისური, ფრანგული, ესპანური და ა. შ., მაგრამ არც ერთს არ აქვს ქართულის მხარდაჭერა, რაც ართულებს ხარისხიანად ტექსტებთან მუშაობას, კონკრეტულად ამ დროისთვის შეუძლებელია შემდეგი ტიპის ამოცანების გადაჭრა:

Lemmatizating and Stemming - სტემინგის შემთხვევაში გვაქვს წინასწარ აღწერილი ენის წესები და ამ ცოდნის გამოყენებით სისტემა ტოვებს სიტყვის ფუძეს, ხოლო ლემატიზაციის დროს, წინასწარ არსებობს ლექსიკონი, საიდანაც სისტემა იღებს ინფორმაციას და სიტყვა გადაჰყავს საწყის ფორმაში. მაგალითად, ინგლისური სიტყვისთვის WENT(წავიდა) სტემინგი ვერ გააკეთებს ნორმალიზაციას და საწყის ფორმამდე გადაყვანას GO(წასვლა), ხოლო ლემატიზატორი კი დაგვიბრუნებს GO.

Auto-generating Syntax Tree and Parsing - სინტაქსური ხის გენერირება და პარსინგი გვაძლევს ტექსტის წინადადებად დაყოფის, სინტაქსური ანალიზის და სინტაქსური ხის გენერირების საშუალებას.

Named Entity Recognition - სახელთა რაობის ავტომატური ამოცნობა, მაგალითად ქალაქების, ქვეყნების სახელების და ა. შ. ამოცნობა. ერთ-ერთი გამოყენება NEC-ის პრაქტიკაში არის რაობების შეცვლა ტეგებით ტექსტის წინასწარი დამუშავების დროს, რომელიც გააუმჯობესებს კლასიფიკაციის ან კლასტერიზაციის ხარისხს.

Automatic Word-sense Disambiguation - ტექსტებში ომონიმიის ავტომატური მოხსნა. მაგალითად სიტყვა “და” ქართულ ენაში, კონტექსტიდან გამომდინარე შეიძლება იყოს როგორც არსებითი სახელი, ასევე კავშირი.

Spell Checker - ინსტრუმენტი, რომელსაც ავტომატურად შეუძლია ტექსტში ისეთი სიტყვების ფიქსირება, რომლებიც შეიცავენ შეცდომებს და სწორი ვარიანტის შემოთავაზება.

სამწუხაროდ, ამჟამად არ არსებობს ქართული ენისთვის ზემოთ აღწერილი არცერთი ინსტრუმენტი ცალკე, და არც მთლიანად ერთ პაკეტში, როგორც გვაქვს ინგლისურის და სხვა ენების შემთხვევაში. ამ ინსტრუმენტების გარეშე შეუძლებელია პროფესიონალური აპლიკაციების შექმნა სადაც აუცილებელია ნატურალური ენის დამუშავება და გაგება.