UDPipe მოდელის სატესტო შეფასება ქართული ენის სინტაქსურ ხეთა ბანკის გამოყენებით

მარიამ აბრამიშვილი

ნაშრომი წარდგენილია ილიას სახელმწიფო უნივერსიტეტის "მეცნიერება და ხელოვნების" ფაკულტეტზე "გამოყენებითი ენათმეცნიერების" მაგისტრის აკადემიური ხარისხის მინიჭების მოთხოვნის შესაბამისად

სამეცნიერო ხელმძღვანელი: ირინა ლობჟანიძე, დოქტორი

ილიას სახელმწიფო უნივერსიტეტი

თბილისი, 2025

განაცხადი

„როგორც წარდგენილი სამაგისტრო ნაშრომის ავტორი ვაცხადებ, რომ ნაშრომი წარმოადგენს ჩემს ორიგინალურ ნამუშევარს და არ შეიცავს სხვა ავტორების მიერ აქამდე გამოქვეყნებულ, გამოსაქვეყნებელ, მიღებულ ან დასაცავად წარდგენილ მასალებს, რომლებიც ნაშრომში არ არის მოხსენიებული ან ციტირებული სათანადო წესების შესაბამისად.“

მარიამ აბრამიშვილი

2025 წელი

მადლობა

მადლობას ვუხდი ჩემს ხელმძღვანელს, ქალბატონ ირინა ლობჟანიძეს, გაწეული შრომისა და მხარდაჭერისთვის. მისგან მიღებული დეტალური უკუკავშირი და რეკომენდაცია ფასდაუდებელი აღმოჩნდა ნაშრომის წერის პროცესში.

მადლობა ჩემს მშობლებსა და დას, სწავლაში ხელშეწყობისთვის, მაქსიმალური თანადგომისა და ტექნიკური უზრუნველყოფისთვის. მათი მზრუნველობა მუდმივად მეხმარებოდა სირთულეების გადალახვაში.

ამ მხარდაჭერის გარეშე ნაშრომის სისრულეში მოყვანა შეუძლებელი იქნებოდა.

აბსტრაქტი

UDPipe არის სამანქანო დასწავლის მოდული, რომელიც მნიშვნელოვან როლს ასრულებს ტექსტის ტოკენიზაციაში, ტეგირებაში, ლემატიზაციასა და სინტაქსურ დამოკიდებულებათა ანალიზში. აღნიშნული ნაშრომის ფარგლებში მოდული გამოყენებულია ქართული ენის სინტაქსური ანოტირების შესაქმნელად და გაწრთვნილია უნივერსალური დამოკიდებულებების (უდ) სქემის შესაბამისად. ამასთან, უდ ეფუძნება ჩომსკის უნივერსალური გრამატიკის თეორიას და წარმოადგენს მრავალენოვანი პარსერების შემუშავების შესაძლებლობას. მიუხედავად იმისა, რომ მოდული წარმატებით გამოიყენება ასზე მეტი ენისთვის, ქართულ ენაზე კვლევა მხოლოდ საწყის ეტაპზეა.

ნაშრომის მიზანია UDPipe მოდულის გაწრთვნა CoNLL-U ფორმატში დამუშავებულ ქართული ენის სინტაქსურ ხეთა ბანკზე. ნაშრომი იყოფა ხუთ ნაწილად: პირველი ნაწილი მიმოიხილავს უნივერსალური დამოკიდებულებების შექმნის ისტორიასა და პრინციპებს; მეორე ნაწილი დეტალურად განიხილავს UDPipe მოდულის შესაძლებლობებს ქართული ენის ხეთა ბანკის კონტექსტში; მესამე ნაწილი აღწერს ქართული ენის მორფოლოგიურ და გრამატიკულ მახასიათებლებს; შემდგომ, მოცემულია UDPipe მოდელის გაწრთვნის პროცესი ქართულ ენობრივ მონაცემებზე; ნაშრომის ბოლოს კი წარმოდგენილია ლაზური ენის მოდელის შექმნის გამოცდილება და შედარება ქართულ ენასთან.

კვლევა ეფუძნება CoNLL-U ფორმატში არსებულ ქართულ ენობრივ მონაცემთა გაწრთვნას UDPipe მოდულზე. მოდული დამუშავდა ქართული ენის სინტაქსური ხეთა ბანკის გამოყენებით, რაც მოიცავდა მრავალსიტყვიანი ტოკენების, მორფო-სინტაქსური მარკირებისა და სინტაქსური დამოკიდებულებების ანალიზს. ლაზური ენის მოდელის ტრენინგი გამოყენებულ იქნება დამატებით მონაცემთა წყაროდ, რამაც შესაძლებელი გახადა შედეგების შედარებითი ანალიზი.

კვლევის შედეგები აჩვენებს, რომ მოდული განსაკუთრებით ეფექტურია სიტყვის ტოკენიზაციაში, თუმცა დასახვეწია მორფო-სინტაქსური ანოტირება. რეკომენდებულია მრავალენოვანი ანოტირებული მონაცემების გამოყენება, სხვა ქართველური ენების მონაცემთა ინტეგრაცია და ჰიბრიდული მოდელების დანერგვა. კვლევა ხელს უწყობს

ქართული ენის კორპუსის განვითარებას და მის ინტეგრირებას საერთაშორისო კვლევებში.

საძიებო სიტყვები: *ციფრული ჰუმანიტარია; ტეგირება; პარსინგი; UDPipe მოდელი; CoNLL-U.*

Abstract

UDPipe is a machine learning module which is worked for text tokenization, tagging, lemmatization and syntactic dependency analysis. Within this work it has been trained according to the Universal Dependencies (UD) framework and used for the syntactic annotation of Georgian. UD is based on Chomsky's theory of universal grammar and provides the possibility to develop multilingual parsers. Although the module has been effectively applied across more than a hundred languages, research on the Georgian language is still in early stages.

The primary objective of this thesis is to train the UDPipe model on the Georgian syntactic treebank prepared in CoNLL-U format. The study is divided into five sections: the first section revies the history and principles behind the UD; the second section explores the capabilities of the UDPipe model in the context of the Georgian treebank; the third section describes the morphological and grammatical characteristics of the Geirgian language; the forth secrion represents the process of training the UDPipe model on Georgian linguistic data; and the final part discusses the experience of developing a model for the Laz language and compares it to the Georgian language.

The research methodology involves training the UDPipe model using Georgian linguistic data in the CoNLL-U format. The model was developed based on the Georgian syntactic treebank, focusing on the analysis of multi-word tokens, morpho-syntactic annotation and syntactic dependencies. The training of the Laz language model served as an additional data source, enabling a comperative analysis of the results.

The finding indicate that the model performs particularly well in word tokenization but morpho-syntactic annotation requires further redinement. It is recommended to utilize multilingual annotated datasets, integrate data from other Kartveliann languages, and adopt hybrid models. The study contributes to the development of the Georgian language corpus and its intergration ino international research initiatives.

Keywords: *Digital Humanities; Tagging; Parsing; UDPipe; CoNLL-U.*