EUSKARAZKO EGITURA SINTAKTIKO KONPLEXUEN ANALISIRAKO ETA TESTUEN SINPLIFIKAZIO AUTOMATIKORAKO PROPOSAMENA / READABILITY ASSESSMENT AND AUTOMATIC TEXT SIMPLIFICATION. THE ANALYSIS OF BASQUE COMPLEX STRUCTURES

Tesiaren egilea: Itziar Gonzalez-Dios
Unibertsitatea: Euskal Herriko Unibertsitatea (UPV/EHU)
Saila: Euskal Hizkuntza eta Komunikazioa
Tesi-zuzendaria: Arantza Díaz de Ilarraza eta María Jesús Aranzabe
Tesiaren laburpena:

Gure gizartean milaka dokumentu sortzen dira egunero, baina horietariko asko konplexuegiak direnez ez dira guztiontzat eskuragarriak. Testu horiek eskuragarriagoak egiteko, tesi-lan honetan euskarazko testuen konplexutasuna aztertzeko eta testu konplexuen sinplifikazioa automatikoki gauzatzeko lehen urratsak egin ditugu.

Testuen konplexutasunaren azterketari dagokionez, testu bat konplexutzat hartzeko irizpideak finkatu eta egitura konplexuak definitu ditugu euskarazko corpusetan egindako analisietan eta erdal hizkuntzetako lanetan oinarrituta. Horretaz gain, testuak automatikoki sinpleak ala konplexuak diren jakiteko, hainbat mailatako 94 ezaugarri linguistikotan oinarritzen den eta euskarri bektoredun makinak (SVM) sailkatzaile bezala erabiltzen dituen ErreXail sistema sortu dugu.

Konplexutasuna tratatzeko aukeratu dugun bidea Testuen Sinplifikazio Automatikoa (TSA) izan da. Horretarako, testuak sinplifikatzen dituen EuTS sistemaren diseinu linguistikoa egin dugu. EuTS sistemak konplexutasunaren corpus-azterketa linguistikoan oinarritutako erregelak aplikatzen ditu. Sistemak bi sinplifikazio mota (ordezkapen sintaktikoen sinplifikazioa eta sinplifikazio sintaktikoa) egiten ditu eta testuak hiru mailetara (azaleko sinplifikazio sintaktikoa, sinplifikazio naturala eta sinplifikazio absolutua) egokitzen ditu. Hori egiteko bost eragiketa definitu ditugu: ordezkapen sintaktikoen sinplifikazioan, i) azaleko ordezkapen sintaktikoak eta sinplifikazio sintaktikoan, i) banaketa, ii) esaldien berreraikitzea, iii) esaldien ordenatzea eta iv) esaldien zuzenketa eta egokitzapena. Kasu-azterketa bezala, informazio biografikoa duten egitura parentetikoak sinplifikatzen dituen Biografix tresna eleaniztuna inplementatu dugu.

TSAko gure hurbilpena kontrastatzeko, ETSC corpusa osatu dugu. Bertan jatorrizko 227 esaldiren eskuz sinplifikatutako bi bertsio bildu ditugu: estrukturala eta

intuitiboa. Bertsio estrukturala zehaztu ditugun gidalerro batzuei jarraituta lortu dugu eta intuitiboa baten intuizioan oinarrituta. Horiek analizatzeko, etiketatze-eskema bat garatu dugu eta etiketatze-eskema horretan oinarrituta, testuak eskuz sinplifikatzean egin diren eragiketak aztertu ditugu. Analisi hori bi hurbilpenak konparatzeko eta amankomunean dituzten eragiketak lortzeko ere baliatu dugu.

Halaber, konplexutasunaren azterketa automatikoa eta sinplifikazio automatikoa egin ahal izateko, Mugak eta Aposizioak izeneko oinarrizko tresnak sortu ditugu. Informazio linguistikoan oinarrituta, Mugak tresnak perpausen mugak identifikatzen ditu eta Aposizioak tresnak aposizioak eta aposizio-sintagmak identifikatu eta sailkatzen ditu.