EUSKARAZKO ENTITATE-IZENAK: IDENTIFIKAZIOA, SAILKAPENA, ITZULPENA ETA DESANBIGUAZIOA

Tesiaren egilea: Izaskun Fernandez Gonzalez
Unibertsitatea: Euskal Herriko Unibertsitatea (UPV/EHU)
Saila: Lengoaia eta Sistema Informatikoak Saila
Tesi-zuzendaria: Iñaki Alegria Loinaz eta Nerea Ezeiza Ramos
Tesiaren laburpena:

Kalifornian jaio eta non hil zen Walt Disney? Gizakiok, gure ezagutzagatik, badakigu galdera horretan *Walt Disney* espresioak pertsona-izen bati egiten diola erreferentzia eta *Kalifornia*k, aldiz, toki-izen bati. Baina, nola egin makina batek, galdera bat emanik, horrelako espresioak erauzteko (identifikazioa eta sailkapena)? Nola jakin dezake makina batek *Walt Disney* espresioak pertsona bati egiten diola erreferentzia, eta ez pertsona horrek sortu zuen eta izen bera duen erakundeari (desanbiguazioa)? Eta, azkenik, nola bila ditzake makina batek erantzunak beste hizkuntza bateko informazio sorta batean (itzulpena)? Alegia, nola automatizatu daiteke entitate-izen bezala ezagutzen diren espresioen tratamendua? Hizkuntza-prozesamenduaren arloak, urte askotan zehar, arazo horiek eta beste hainbat ebazteko aurrerapauso garrantzitsuak ematen dihardu. Eta, euskararen munduan, IXA taldea horretan lan handia egiten diharduen taldea da. Ixa taldean landutako tesi honek hain zuzen ere entitate-izenak automatikoki lantzea du helburu. Bereziki hiru alor nagusi landu dira:

- *Euskarazko entitate-izenen identifikazio eta sailkapena:* euskarazko testuetan agertzen diren entitate-izenak automatikoki identifikatu eta sailkatzeko tresnaren garapena.
- *Euskarazko entitate-izenen itzulpena:* itzulpen automatikorako zein galdera-erantzunen aplikazio eleaniztasunetarako oso lagungarri gertatzen diren entitate-izenen aipamen elianiztunak automatikoki sortzeko estrategia da eginkizun honen funtsa.
- *Euskarazko entitate-izenen desanbiguazioa:* euskarazko testuetan agertzen diren entitate-izenen agerpen anbiguoak automatikoki desanbiguatzeko tresna garatzea.

Horietarako guztietarako, metodologikoki hiru irizpide komun izan dira:

- Euskara baliabide urriko hizkuntza izanik, baliabideak berrerabitzea, eta metodo ez-gainbegiratu edota erdi-gainbegiratuei lehentasuna ematea.
- Hizkuntza-ezagutzan oinarritutako eta ikasketa automatikoko teknikak entitate-izenen atazak ebazteko ustiatzea eta ahal denean konbinatzea, metodo sofistikatuen erabilera ekidinez.
- Entitate-izenen atazak automatizatzean, euskararen ezaugarri morfosintaktikoek duten eragina aztertzea.