LATENT SEMANTIC INDEXING ETA IKASKETA AUTOMATIKOA HIZKUNTZAREN PROZESAMENDUAREN ARLOAN: TESTU-SAILKATZEA, HITZEN ADIERA-DESANBIGUATZEA ETA KORREFERENTZIA-EBAZTEA SVD BIDEZKO DIMENTSIO-MURRIZKETA ETA MULTI-SAILKATZAILEA KONBINATUZ

Tesiaren egilea: Ana Zelaia Jauregi
Unibertsitatea: Euskal Herriko Unibertsitatea (UPV/EHU)
Saila: Konputazio Zientzia eta Adimen Artifiziala
Tesi-zuzendaria: Olatz Arregi Uriarte eta Basilio Sierra Araujo
Tesiaren laburpena:

Latent Semantic Indexing (LSI) testuen semantika jasotzeko gaitasuna duen tresna bat da. Oinarri matematikoa du, eta bi teknika konbinatzen ditu: aljebra linealeko bektore-espazioak, batetik, eta matrize-deskonposaketarako Singular Value Decomposition (SVD) metodoa, bestetik. Testu multzo (corpus) batetik abiatuta sortzen da espazio semantikoa eta bertan kokatzen dira hitzak eta testu zatiak adierazten dituzten bektoreak. Espazioaren dimentsioa murriztean hitzen eta testu zatien arteko erlazio semantikoak hobeto erakustea lortzen da. Bektoreen arteko angeluaren kosinua erabiltzen da haien arteko konparaketa semantikoak egiteko; horrela, hitzek edo testuek duten antzekotasun semantikoa neur daiteke.

Konputagailuek ikasteko gaitasuna garatzea helburu duen informatikaren alorra da Ikasketa Automatikoa, makinak ere giza adituek erakusten duten trebeziarekin problemak ebazteko eta erabakiak hartzeko gai izan daitezen. Ikasketa automatikoko metodoak azken urteotan Hizkuntzaren Prozesamenduaren hainbat atazatan oso lagungarriak gertatzen ari dira.

Ikerketa-lan honetan LSI eta Ikasketa Automatikoa uztartzeak Hizkuntzaren Prozesamendu Automatikoaren hainbat atazaren ebazpenean ekar dezakeen onura aztertzen da. Izaera oso ezberdineko hiru atazarekin egin dugu lan:

- **Testu Sailkatzea**. LSIren aplikazio-eremu tradizionala da. Oso izaera desberdineko testuen sailkatzearekin probak egin eta bi hizkuntzarekin esperimentatu dugu: euskarazko eta ingelesezko testuekin. Esperimentu bakoitzean, testu-dokumentu guztiak batera hartuz osatu da LSIrako corpusa, informazio orokorra biltzen duen corpusarekin esperimentatzeko asmoz. Sailkatze-problema etiketa anitzetarako (*multi-labeling*) estrategia bat diseinatu da.

- **Hitzen Adiera Desanbiguatzea**. Polisemikoak diren hitz-adierak desanbiguatzeko, hitzen agerpenen testuinguruaz baliatzea erabaki dugu. Nazioarte mailako txapelketa batean parte-hartzea aukera aparta izan da metodoaren eraginkortasuna neurtzeko. Ataza honetarako ezagutza espezializatua jasoko duten corpus txiki askorekin esperimentatzea erabaki dugu.
- **Korreferentzia Ebaztea**. Entitate berari erreferentzia egiten dioten diskurtsoko agerpenak haien artean erlazionatzea lan zaila, baina aldi berean garrantzitsua da testuaren ulermena eskatzen duten atazetan. Informazio lexiko, morfologiko eta sintaktikoaz gain, beharrezkoa gertatzen da informazio semantikoa eta pragmatikoa erabiltzea korreferentzia ebazteko. Ataza honen ebazpenerako LSI aplikatzea ez da berehalakoa gertatzen. Oraingoan ere euskarazko eta ingelesezko corpusekin egin dugu lan.

Egindako lanaren ekarpen nagusiak hiru dira: (1) Erabilitako metodologia, LSI+Ikasketa Automatikoa, oinarrizko sailkatzaileekin eta multi-sailkatzaileekin, (2) LSIren aplikazio-eremu ez hain berehalakoetan haren portaera aztertzea, etorkizunean bide berriak urratzeko metodologiak eman lezakeena aztertuz, eta (3) metodologia hori, euskarazko corpusak baliatuta, Hizkuntzaren Prozesamenduko hiru atazetan frogatzea.