Baliabide lexiko-semantikoak hizkuntz industriarako

Jose Maria Arriola / Arantza Diaz de Ilarraza / Kepa Sarasola

Hizkuntzaren inguruan industria berria sortzen ari da, konputagailuaz baliaturik hizkuntza tratatzea helburu duena. Arlo honek aurrera egin dezan, ezinbestekoak dira hitzen esanahia emango duten baliabide lexikalak. Europako Batasuneko *Language Engineering* programaren irizpideetan baliabide lexikalek duten oinarrizko papera azpimarratuta azaltzen da.

Mintzagai dugun Hiztegia 2002 proiektuak ere Europako Elkartearen laguntza du (FEDER, 2FD97-2000-2001) eta baliabide lexikalak sortzea jomuga izan duten honako proiektu hauekin dago lotuta: Wordnet[2], EuroWortdnet eta ITEM[3]. Proiektu honen bidez, IXA[4] taldeak ondorengo baliabide lexikalak erdietsi nahi ditu:

– Euskal Hiztegiaren bertsio egituratua; horretarako, TEIko (*Text Encoding Initiative*) gida-lerroei jarraiki. SGML (*Standard Generalized Markup Language*) lengoaia estandarra erabiliko da.
– Euskal Hiztegiaren ezagutza-base lexikala: bertatik ateratako erlazio semantikoez osatua.

– Euskal Wordnet: EuroWordnet egokitzea, ingelesezko kontzeptuei euskarazkoak lotuz.

Baliabide horiek sortzean, besteak beste, ondorengo produktu komertzialak garatzea dugu helburu:

– Euskal Hiztegiaren bertsio elektroniko egituratua (CD-ROMean, Interneten edota testu-prozesadoreetan integratua)
– Euskararako *thesaurus* bat testu-prozesadoreetan integratua: sinonimia, hiperonimia, hiponimia eta bestelako kontzeptuen erlazioak kontsultatu ahal izateko.

Historikoki, baliabide lexikalak eskuz egiten ziren; baina, informazio-kopuru itzela landu behar zela-eta ahalegin handia eskatzen zuela kontuan izanik, laguntza automatiko edo erdiautomatikoen bidea jorratzeari ekin zaio azken hamarkadan. Hiztegietan eta corpusetan dagoen informazioa aterata ezagutza-base lexikalak (EBL) egin izan dira. EBLak baliabide lexikal egituratuak ditugu, hitz eta adierei buruzko informazioa dutenak, alegia. Esaterako, dohainik banatzen den Wordnet EBLan ingelesezko adiera bakoitza hitz sinonimo multzo batekin (*synset* deritzonarekin) adierazten da eta adiera guztiak hierarkikoki antolatuta daude. EuroWordnet diseinu bereko beste EBL bat da eta Europako zortzi hizkuntzatara (aleman, espainiera, estoniera, frantses, ingeles, italiera, nederlandera eta txekierara) zabaldu da.

EBL gehienak ingeleserako sortu direnez, gainontzeko hizkuntzak teknologia berriekiko egoera ahulean daude. Egoera horri aurre egiteko bi irtenbide osagarri ikusten ditugu:

1. Hizkuntza bakoitzarentzat dauden corpus eta hiztegietatik abiatuta EBLak sortzea. Gure kasuan, iturri lexikal gisa. Euskal Hiztegia baliatu dugu. Lehenbiziko lana Euskal Hiztegia SGML-TEI estandarrei jarraituz egituratzea izan da. Horrela bada, euskara aztergai edo lan-tresna duen edonorentzat baliagarri izango da. Bertsio egituratu honen definizioak aztertuz, hainbat erlazio lexiko-semantiko aterako ditugu: sinonimia, hiperonimia (*klase-azpiklase* erlazioa; adibidez: animalia-intsektu), meronimia (*osoa-partea* erlazioa; adibidez: txori-moko), etab.

2. Beste hizkuntzetarako EBLak sortzeko, ingeleserako egin diren EBLez baliatzea. Gure kasuan, EuroWordnet abiapuntutzat hartuz euskararako Wordnet-a egin nahi dugu, ingelesezko kontzeptuei euskarazkoak lotuz. Euskararako Wordnet hori egiteko, metodo erdiautomatikoak erabiliko ditugu, baina gero emaitzak eskuz orraztuko ditugu.

Irudian ikus daiteke EuroWordnet-i euskarazko kontzeptuak lotzeko erabiltzen dugun interfazea[5].

Proiektuaren izenburua:

HIZTEGIA 2002: Hizkuntz teknologiako aplikazioetarako ezagutza-base lexiko-semantikoa

Helburua:

Euskararen ontologia eraikitzea, hau da, euskarazko hitzen esanahiak modu hierarkikoan definitzea eta beren arteko hainbat erlazio lexiko-semantikorekin osatzea (sinonimia, hiperonimia, meronimia…). Horretarako, Ibon Sarasolaren Euskal Hiztegia eta EuroWordnet[1] ontologia izango ditugu oinarri.

Bi azpihelburu ditugu: alde batetik, Euskal Hiztegiko adieren definizioetatik erlazio lexiko-semantikoak ateratzea eta, bestetik, EuroWordnet-eko ingelesezko kontzeptuei euskarazko kontzeptuak lotzea.

Zuzendaria:

Arantza Díaz de Ilarraza.

Ikerketa-taldea:

Lan-taldea: Agirre E., Ansa O., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., García E., Irurtzun A., Lersundi M., Martinez D., Pociello E., Sarasola K., Soroa A., Urkia M., Zubelzu E.

Partaideak: UPV-EHUko IXA taldea, UZEI eta Ibon Sarasola

Saila:

Lengoaia eta Sistema Informatikoak

Fakultatea:

UPV-EHUko Informatika (Donostia)

[1] http://www.let.uva.nl/~ewn/
[2] http://www.cogsci.princeton.edu/~wn
[3] http://sensei.ieec.uned.es/item/
[4] http://ixa.si.ehu.es/
[5] EuroWordnet kontsultatu nahi duenak jo beza ondorengo helbidera: http://ixa.si.ehu.es/wei3.html