Corpus-kontsultan hitzen agerpenak antzekotasun semantikoaren arabera multzokatzea (*sense clustering*)

Proposer(s) / Proposatzailea(k):

Elhuyar Hizkuntza Teknologiak

Antton Gurrutxaga, PhD

Hizkuntza Baliabideak eta Tresnak zerbitzu-lerroko ikertzailea

Contact / Kontaktua:

a.gurrutxaga@elhuyar.eus

Tel.: 943363040 | luzp.: 220 I mugikorra: 655716254

Zelai Haundi, 3

Osinalde industrialdea

20170 Usurbil

Description / Deskribapena

Hiztegigintza erdiautomatikoaren arloa garrantzia hartuz joan da azken hamarkadetan. Batetik, hizkuntzaren prozesamendu automatikoak testuetatik gero eta informazio ugariagoa eta aberatsagoa eskuratzeko aukera eman du. Hasieran hitzen informazio estatistiko oinarrizkoa zena (agerpen-maiztasunak) sofistikatuz joan da. Hiztegigintzari berari begira, lexikografoen eta terminologoen ohiko egitekoak dira automatizatu nahi direnak; hain zuzen ere: patroi morfosintaktiko eta kolokazionalen erauzketa, unitate fraseologikoen eta terminoen erauzketa, adibide egokien hautaketa, definizio-erauzketa, eta adiera-bereizketa. Euskaraz, egiteko horietako batzuk automatizatzeko teknologia garatu da, edo garatzeko bidean da, behintzat. Une honetan, adiera-bereizketa da esfortzua egitea eskatzen diguna.

Adiera-bereizketa hiztegigilearen lan zailenetakoa da. Corpuseko agerpenak (edo horien multzo bat) aztertuz, edo introspekzioa (*insight*) erabiliz, lexikografoek adieratan banatzen dute hitzaren semantika, dela intuitiboki, dela irizpide edo metodologia batean oinarrituta. Metodologia horren funtsezko ideia izaten da hitzen esanahiak edo adierak diskretuak eta bereizteko modukoak direla.

Proiektu honetan planteatzen dugun egitekoa da lexikografoaren disekzionatze-lan horri laguntzea, era honetara: corpus-datu kopuru handiak prozesatuz eta analizatuz, hitz jakin baten adierak bereizteko materiala eskaintzea. Material hori corpus-agerpenak antzekotasun semantikoaren araberako multzoak edo klusterrak dira. Horrela definituta, ataza gainbegiratu gabeko *sense clustering* erakoa da. Bestetik, material hori bera baliagarria izango da corpusen kontsulta-sistemak hobetzeko, erabiltzaileari hitz baten agerpen-multzo osoa eta gordina erakutsi beharrean, automatikoki bereizitako adiera-klusterren agerpen "egokienak" eskaintzen bazaizkio. Elhuyarren, Ber2tek proiektuaren

testuinguruan, adibide egokiak aukeratzeko oinarrizko teknika batzuek garatu ditugu, eta horiek aplika litezke multzokatze-prozesuaren irteera dosifikatzeko.

Klusterrak antzekotasun semantikoaren arabera bereizteak berekin dakar antzekotasun distribuzionala konputatzeko teknikak erabiltzea. Azken urteotan, *word embeddings* eta sare neuronaletan oinarritutako metodo konputazional berriek aurrerapen handia ekarri dute arlo honetara. Ideia da, beraz, teknika horiek euskararen hiztegigintza automatizatzeko ataza honetan aplikatzea.

Goals / Helburuak

Proiektuaren helburu nagusia euskarazko hitzen corpus-agerpenak antzekotasunaren semantikoaren arabera multzokatzea, eta hiztegigilearen lan-interfazean informazio hori osorik ez ezik dosifikatuta ere ikusteko aukera ematea, kluster bakoitzetik agerpen "egokienak" adibidetzat hautatuta.

Requirements / Betebeharrak

- Euskara
- Python
- Antzekotasun semantikoa kalkulatzeko metodoak
    - *word embeddings* (word2vec, Gensim…)
- Multzokatze- edo clustering-teknikak

Framework / Esparrua

Adiera-bereizketa erdiautomatikoa Elhuyar Hizkuntza eta Teknologia unitateko zenbait zerbitzu eta produktutan integratzeko asmoa dugu. Zehazki, corpus-kontsultarako sistema aurreratuetan, patroi morfosintaktiko/kolokazionalekin eta adibide egokiak hautatzeko sistemarekin batera.

Tasks and plan / Atazak eta plana

- 1. Artearen egoera aztertzea
    - Antzekotasun semantikoaren kalkulua eta haren aplikazioa ataza honetan
- 2. Diseinu esperimentala
    - Hitz baten agerpenak semantikoki multzokatzeko metodoa diseinatzea
- 3. Esperimentazioa eta ebaluazioa
- 4. Prototipoaren garapena