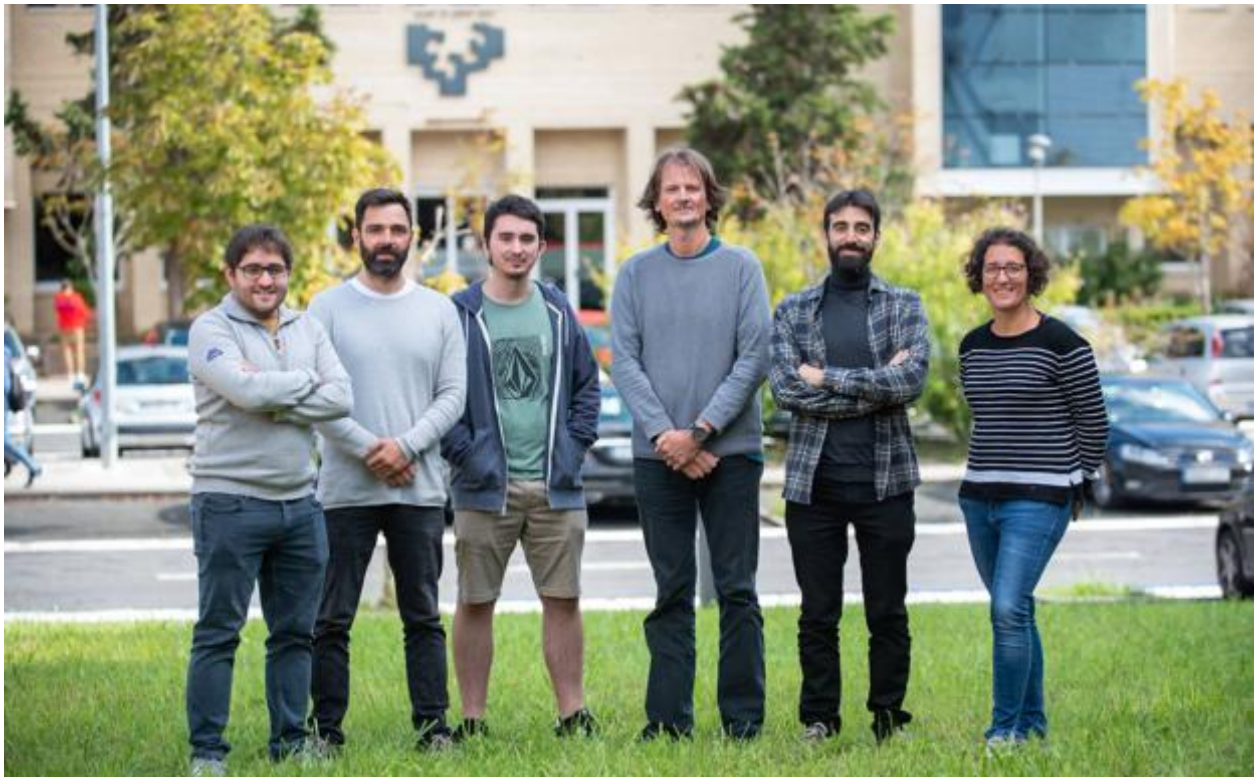
www.diariovasco.com /sociedad/hizkuntza-teknologiak-eraldatzeko-20221006123131-nt.html

Eredu neuronal berriak hizkuntza-teknologiak eraldatzeko

El Diario Vasco ⋮⋮ 06/10/2022

Ezkerretik eskubira, Gorka Urbizu (Orai), Xabier Saralegi (Orai), Ander Salaberria (Hitz zentroa), Aitor Soroa (Hitz zentroa), Aitor García-Pablos (Vicomtech) eta Montse Cuadros (Vicomtech).



'DeepText' proiektuaren baitan egin diren aurrerapenek «bultzada nabarmena» emango diete euskararen eta gaztelaniaren hizkuntza prozesamenduan oinarritutako aplikazioei

Jueves, 6 de octubre 2022, 13:01 | Actualizado 16:12h.

Adimen artifizialeko azken teknikak erabiltzen dituzten hizkuntza eredu neuronalak sortu dituzte Euskal Herriko Unibertsitateko HiTZ zentroak eta Orai NLP eta Vicomtech zentro teknologikoek 'DeepText' proiektuaren baitan. Ohar baten bidez adierazi dutenez, lortutako aurrerapenek «bultzada nabarmena»emango diete euskararen eta gaztelaniaren hizkuntza prozesamenduanoinarritutako aplikazioei.

Bi urtez aritu dira HiTZ, Orai NLP eta Vicomtech zentroetrako ikertzaileak 'DeepText' proiektuan lanean, HiTZ zentroa buru dela. Adimen artifizialeko hizkuntza eredu neuronalen belaunaldi berria sortzea izan dute helburu, Euskal Herriko industriaren hizkuntza teknologiak eraldatzeko.

Izan ere, ekoizpen zientifikoak eta garapen teknologikoak, oro har, ez dute kontuan hartu gaztelania ingelesa bezain beste, eta are gutxiago euskara. Horren ondorioz, orain arte ez da aukera handirik izan hizkuntza naturalaren prozesamendua eta horri lotutako zerbitzuak garatuz hizkuntza teknologietako eta adimen artifizialeko sektorea eraldatzeko.

Azken bekaunaldiko ereduak

Arlo honetan euskarak eta gaztelaniak duten egoera hobetzeko, bi hizkuntzotarako azken belaunaldiko hizkuntza eredu neuronalak sortu dituzte (euskararako, lehenak), baita hizkuntza eredu neuronal eleaniztunak ere (euskara, gaztelania, frantsesa eta ingelesa biltzen dituztenak).

«Hizkuntza naturalaren prozesamenduaren helburua da makinak gure hizkuntza ulertzeko eta sortzeko gai izatea, horri esker zenbait ataza egiteko ahalmena izateko», azaldu dute partzuergoko ikertzaileek.

Orain arte horretarako erabili izan diren teknikak zaharkituta geratu dira, eta hizkuntza eredu neuronaletan oinarritutako sistemak erabiltzen dira orain. Azken urteetan, paradigma aldaketa erabat disruptiboa gertatzen ari da hizkuntza naturalaren prozesamenduan.

«Hizkuntza eredu neuronal generikoak entrenatzen dira testu corpus erraldoiak erabiliz, hizkuntzaren ezagutza orokor bat izan dezaten, eta, gero, doitu egiten dira ataza jakin bat egiteko gai izan daitezen (bilaketak egin, testuen gaiak sailkatu, testuetako sentimenduak detektatu, laburpen automatikoak egin eta abar)», zehaztu dute.

Euskarazko corpusik handiena

Baliabide urriko hizkuntzek arazoak dituzte halako corpus handiak osatzeko, baina 'DeepText' proiektuan euskararako inoiz izan den corpusik handiena osatu da: 350 milioi hitzeko corpusa. Hala, corpus hori eta IXA Taldeak sortutako 288 milioi hitzeko euscrawl corpusa erabilita, euskararako lehenengo hizkuntza eredu neuronalak sortu dituzte, paradigma berria erabilita, eta hainbat ataza egiteko entrenatu dituzte, sistema berrietan ezartzeko.

Hizkuntza eredu neuronal eleaniztunak baliabide urriko hizkuntzetarako tresnak ezartzeko erabiltzen dira. Azaldu dutenez, «munduan 7.000 hizkuntza inguru daude, gehienak baliabide urrikoak. Corpus eta material digital gutxi dutenez, zailtasunak dituzte entrenamendu adibideak sortzeko. Euskara ere multzo horretan sar dezakegu. Halakoetan, hizkuntza eredu eleaniztunak erabiltzea alternatiba eraginkorra da. Oinarri hori hizkuntza handi bateko adibideekin entrenatzen da, eta gero euskarazko datuekin probatzen da ea zer emaitza ematen dituen ikusteko».

Ikertzaileek aitortu dute 'transfer learning' deritzon teknikak ez dituela emaitza «perfektuak» ematen, baina adierazi dute «oso emaitza interesgarriak» ematen dituela, adibidez, galdera-erantzun bidezko bilaketak egiteko.

Horrekin batera, ebaluazio ingurune bat sortu dute, hizkuntza eredu neuronalek hizkuntza ulertzeko zenbaterainoko gaitasuna duten neurtzeko eta alorreko ikerketak aurrera eramateko. «Ebaluazio ingurune horrek zenbait ataza linguistiko biltzen ditu (izen berezien detekzioa, sentimenduen detekzioa,

gai sailkapena, korreferentziak ebaztea, galderak erantzutea). Euskara eta gaztelania ebaluatzeko ingurunea sortu dugu», azaldu dute.

Oinarriak jartzen

Ikertzaileek garrantzi berezia eman diote euskara ebaluatzeko atalari (BasqueGLUE), hizkuntza horretarako lehena baita, eta, nabarmendu dutenez, «oso ezinbesteko pauso bat eman dugu Euskal Herriko hizkuntza teknologiak garatzeko bidean».

«Bi urte hauetan, hizkuntza teknologiek aurrera egiteko behar duten oinarri teknologikoa ikertu dugu euskara, gaztelania, ingelesa eta baliabide urriko beste hizkuntza batzuetarako. Gaur egun, hizkuntza teknologietako produktuak garatzeko eta emaitzarik onenak lortzeko beharrezkoak dira hizkuntza eredu neuronalak. Orain arte euskararako horrelako eredurik ez zen sortu. Hizkuntza eredu neuronalak nola erabili aztertu dugu, eta ataza jakinak egiteko doitu, eta, bestalde, hizkuntzen arteko eta domeinuen arteko transferentzia nola egin ikasi dugu», azaldu dute.

Azkenik, gogotsu agertu dira funtsezko ikerketan jarraitzeko, eredu neuronaletan oinarritutako teknika berritzaileak asmatzeko eta horiekin esperimentatzeko, eta espero dute I+G proiektuak bultzatzeko politika publiko eta funtsetan isla izatea.

- Temas
- Zabalik