EusCrawl: kalitate handiko euskal corpus librea

ALBISTEAK 2022/03/25

Rodrigo Agerri eta Aitor Soroa ikerlariak Informatika Fakultatean. Argazkia: Nagore Iraola. UPV/EHU

Adimen artifizialaren erronka nagusietako bat konputagailuek gizakion hizkuntza ulertzea da, eta hori da hain zuzen Hizkuntzaren Prozesamenduaren helburua. Adimen artifizialaren arlo honek iraultza handia jazo du azken urteetan, ikasketa sakona edo "deep learning" teknikei esker eta, zehatzago esateko, hizkuntza-eredu deritzon teknologiari esker.

Hizkuntza-ereduak testu kopuru handiak erabiliz entrenatzen dira, eta, testua irakurriaz, gai dira hizkuntzaren egitura ikasi eta testu berriak sortzeko. Gaur egungo hizkuntzaren prozesamenduko aplikazioen muinean aurki ditzakegu hizkuntza-ereduak, dela bilaketa eta galderen erantzunean, itzulpen automatikoan, ahotsaren ezagutzan edo elkarrizketa-sistema zein txatbotetan. Labur esateko, hizkuntza-ereduak dira hizkuntzaren inguruan egiten diren aplikazio gehienen motorra, eta testuak dira motor horren gasolina.

Hizkuntza-eredu onak eraikitzeko behar den testu kopurua astronomikoa da. Ingelesa bezalako hizkuntzetarako testuak aurkitzea ez da arazoa; nahi adina testu dugu hizkuntza horretan

Interneten. Testu multzo izugarri handiak batu izan dira horrela, adibidez 156 mila miloi hitz dituen Colossal Clean Crawled Corpus (C4) izeneko corpusa. Pertsona batek 2000 urte beharko lituzke hori dena irakurtzeko, egunean 10 ordu irakurriz gero. Horiei lotuta eraikitako hizkuntza-ereduak ere erraldoiak dira, tartean BERTlarge (350 milioi parametro), eta ezagunena, komunikabideetan hainbat aldiz aipatu den GPT-3 (175 mila miloi parametro). Hizkuntza-eredu horiek adimen artifizialean eraiki izan diren gailu konplexuenetakoak dira parametro kopuruan, eta milioika euro gastatu izan dira beraiek entrenatzeko behar den konputazioan (adibidez, 4 miloi dolar inguru GPT-3 entrenatzeko).

Euskara bezalako baliabide urriko hizkuntzetarako, baina, tamaina handiko testu masak biltzea arazo zaila da. Euskararen kasuan existitzen diren eta eskura dauden testu masa handienak Google eta Meta-AI (lehen Facebook) enpresek Internetetik automatikoki jaitsi eta dokumentuen hizkuntza programa bidez identifikatu izan dituzten mC4 eta CC100

corpusak dira. Lehenbizikoak euskarazko mila miloi hitz dauzka eta bigarrenak 416 miloi hitz. Horien kalitatea zalantzan jarri izan da ordea, Internet zaratatsua delako eta dokumentuak euskaraz daudela ziurtatzen duen programa automatikoak akatsak egiten dituelako.

EusCrawl-en garrantziaz

Gabezia horri erantzutera dator EusCrawl. Corpusa osatzen duten dokumentuak modu librean bana daitezke, Creative Commons lizentzia libreekin. 12.5 milioi dokumentu eta 423 milioi hitzez osatuta dago, eta eskuz aukeratutako Interneteko hainbat webgunetatik dokumentuak xurgatuz (crawl ingelesez) osatu da.

Corpusarekin batera, EusCrawl-ekin entrenatutako bi hizkuntza-eredu sortu ditugu, horietako bat egun euskararako dagoen eredurik handiena, 355 Milioi parametrokoa.

EusCrawl corpusa libre izateak euskarak duen nazioarteko ikusgarritasuna areagotzen du, eta mundu zabaleko ikertzaileek euskararako baliabide hobeak sortzea dakar horrek. Esate baterako, dagoeneko badakigu EusCrawl BigScience proiektuan erabiliko dela, helburu bezala hizkuntza-eredu eleaniztun eta erraldoi librea eraikitzea duen proiektua, horretarako bost milioi konputazio-ordu erabiliz. Hortaz, sortutako hizkuntza-ereduak euskaraz ere jakingo du. EusCrawl bezalako baliabideak libre jartzea urrats ezinbestekoa da euskara plaza digitalera jalgi dadin.

Hizkuntzaren prozesamendua eta adimen artifizialaz aparte, EusCrawl corpusa baliabide ezin hobea da hizkuntza bera aztertu nahi duenarentzat. Ez da ahaztu behar corpusen ustiapena dela gaur egun hizkuntzalaritzaren muinetako bat, hizkuntzaren erabilera errealaren gordailuak diren neurrian. Euskarazko corpus handiak bildu izan dira aurretik ere, eta publikoki kontsultagarri jarri, baina EusCrawl osorik deskargatu eta berrerabiltzeko aukera dago. Azpimarratu behar da ez dela gauza bera corpusa kotsultagarri jartzea ala deskargatzeko moduan jartzea. Kontsulta soilek ez dute aukerarik ematen benetako azterketa linguistikoak eta ikerkuntzak egiteko.

EusCrawl-i esker ikasi dugunaz

Corpusa biltzarekin batera, EusCrawl-ekin sortutako hizkuntza-ereduak beste corpusekin sortutakoekin alderatu ditugu, hizkuntzaren prozesamenduko hainbat atazatan beraien kalitatea neurtuaz. Esperimentuek adierazten dute garrantzitsuagoa dela testu kopurua, testuen kalitatea baino. Gaur egun ezagunak diren euskarazko corpus guztiak bilduta ere, hizkuntza nagusien corpusen tamainatik oso urruti geldituko ginateke, eta horrek euskarazko hizkuntzaereduei goi-borne bat ezartzen die. Ondorioz, arriskua dago euskararentzat sor daitezkeen tresnen kalitatea ingelesa bezalako beste hizkuntzen mailara ez iristeko.

Horren aurrean, euskara eta baliabide urriko beste hizkuntzen teknologiak aurrera egin dezan, bi helburu estrategiko azaltzen zaizkigu.

-Corpus handiagoak biltzea, euskaraz ekoizten den eduki gehiago eskuragarri jarriz. EusCrawl eraikitzea posible izan da Berria, Argia, Hitza eta beste hainbat euskal komunikabideei esker, edukia lizentzia librean banatzen dute eta. Ezinbestekoa da gainontzeko ekoizleak ere bide horretara batzea.

-Testu gutxiagorekin ikasiko duten hizkuntza-ereduen ikerketa sustatzea. Tamalez aurreko ahaleginak muga bat du, hizkuntza baten idazten den testu kopuruaren araberakoa. Egun dauden teknikekin eraikitako metodoez haratago, testu gutxiagotik ikasiko duten hizkuntza-ereduak behar ditu euskarak. Euskararako tresnak kalitatezkoak izan daitezen estrategikoa da ikerketa-lerro hau bultzatzea.

Corpusa http://ixa.ehu.eus/euscrawl helbidean aurki daiteke, eta xehetasun guztiak, berriz, https://arxiv.org/abs/2203.08111 artikuluan. EusCrawlekin sortu diren hizkuntza-ereduak zein ikerketa-esperimentuak Hitz Zentroa eta Meta-AI erakundeen arteko elkarlana izan da.

partekatzea