Euskara

EusCrawl, euskarazko hitzen korpus erraldoia

EHUko informatikari euskaldunen IXA taldeak orain arte osatutako euskarazko hitz-corpus handiena bildu du, prozesatu du Hitz zentroaren partaidetzarekin (eta Meta enpresaren laguntzarekin ere), eta berrerabilpenerako prestatu du zenbait formatutan. Creative Commons lizentziekin jarri dituzte erabilgarri materialak, EusCrawl izenarekin.

SUSTATU @sustatu

2022ko martxoaren 24a

Guztira 12.5 milioi dokumentu eta 423 milioi hitzez osatuta dago, eta eskuz aukeratutako Interneteko hainbat webgunetatik dokumentuak xurgatuz (crawl ingelesez)

helbidea: http://ixa.ehu.eus/euscrawl/

Zenbait iturritan dute jatorria testuek, eta horien arabera da berrerabilgarri edukia lizentzia batekin edo bestearekinb: Cc-by-sa lizentzia librearekin eskuratu da edukia Wikipediatik, Berriatik eta Argiatik. Beste murrizketa batzyk dituzte Hitza batzuen edukiek, edo Bilbo Hiria Irratitik eskuratuek.

Zertarako erabili ahal izango da EusCraweleko korpus handi hori? Adimen artifizialean oinarritutako hizkuntza-ereduen teknologian izango du aplikazioa. IXA taldeak azaldu duen bezala, "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 oso handia da. Ingelesa bezalako hizkuntzetarako testuak aurkitzea ez da arazoa; baina hala ere, kopurun horiek bildu egin behar dira, eta horrela zientzialariak lanak hartu dituzte Colossal Clean Crawled Corpus (C4) izeneko corpusa sortzeko aidbidez, 156.000 milioi hitz dituena.

EusCrawl konparazioan, txikia da, baina nonbait hasi behar. Gainera, euskararen kasuan egon dira testu-masa handiak sortuta, baina kalitatearen aldetik ez omen guztiz fidagarriak: Google eta Meta-AI (lehen Facebook) enpresek Internetetik automatikoki jaitsi eta dokumentuen hizkuntza programa bidez identifikatu izan dituzten mC4 (1.000 milioi

Izatez, EusCrawl horiek baino txikiagoa izan arren, erabili dute jada eratorritako beste produktu batzuk sortzeko ere: IXAkoek EusCrawl-ekin entrenatutako bi hizkuntza-eredu sortu dituzte, horietako bat egun euskararako dagoen eredurik handiena, 355 Milioi parametrokoa.

Era berean IXAkoek jakinarazi dute EusCrawl erabiliko delaBigScience proiektuan, helburu bezala hizkuntza-eredu eleaniztun eta erraldoi librea eraikitzea duen proiektua, horretarako bost milioi konputazio-ordu erabiliz. BigScience-ren sortuko den hizkuntza-ereduak euskaraz ere jakingo du.

EusCrawl Interneten argitaratu da, eta IXA taldeko bost lagunek egindako lan gisa ere aurkeztu da, *paper* akademiko batean. EHUko IXA taldearen emaitza dela esan daiteke, baina Meta enpresak ere (Facebook zenak) parte hartu du lanean, IXAn zein Metan zubi egiten duen Mikel Artetxe informatikariaren bidez. Paperra sinatzen dute halaber Itziar Aldabe, Rodrigo Agerri, Olatz Perez de Viñaspre eta Aitor Soroak.

EusCrawli buruzko informazio gehiago, Unibertsitatea.net-en.

Kanal hauetan artxibatua: Euskara | Teknologia berriak