ATRÁS

IÑAKI ALEGRIA, ARANTZA DIAZ DE ILARRAZA, GORKA LABAKA ETA KEPA SARASOLA

Euskarazko Itzulpen Automatikoa garai berrian sartu da

IXA Taldea

CATHEDRA 25/01/2019

Iñaki Alegria, Gorka Labaka, Arantza Diaz de Ilarraza y Kepa Sarasola. Foto: Nagore Iraola. UPV/EHU.

Este artículo se publica en el idioma en que ha sido escrito.

Adimen artifizialean orain dela 25 urte sortzen ziren tresnak, oro har, adituek erregeletan kodetutako jakintzan oinarritzen ziren, mende aldaketak teknika estatistikoen erabileraren nagusitasuna ekarri zuen, eta, azkenaldian, ikasketa sakona (deep learning) eta neurona-sareak (neural networks) lortzen ari dira emaitza hoberenak. Itzulpen automatikoak ere antzeko ibilbidea egin du: hasierako sistemak erregeletan oinarrituta zeuden, mende berriarekin itzultzaile estatistikoak gailendu ziren, eta azken urteetan neurona-sareetan oinarritutako itzultzaileak dira arreta guztia erakartzen dituztenak. Itzultzaile neuronalek lortu dezaketen kalitatea erraz ikus dezakegu Google Translate (https://translate.google.com) edo DeepL (https://www.deepl.com/translator) sistemekin aproba eginda, hori bai, oraingoz gaztelera-ingelesa bezalako hizkuntza handietarako bakarrik eskaintzen dituzte emaitza bikain horiek. Baina, hau da poza, euskararako ere antzeko emaitza onak lortu berri ditugu MODELA proiektuko ikerketarekin.

Azken hamarkadan nagusi izan diren itzultzaile estatistikoak hitzen kontaketa eta probabilitate estimazioetan oinarritzen dira. Lehendik egindako itzulpen bildumak aztertu, hitzak nola itzuli izan diren zenbatu, eta testu berri bat itzultzean kontaketa horien araberako itzulpen probableena sortzen saiatzen dira. Sistema hauek ez dute esaldiaren egiturari buruzko informaziorik, testuinguru lokaletan oinarritzen dira, soilik. Hitz solteak edo jarraian agertzen diren 3 edo 4 hitzetako sekuentziak hartzen dituzte kontuan asko jota. Horrela, ez da harritzekoa sortzen dituzten itzulpenak jariotasun falta izatea eta zatika sortu izanaren itxura ematea. Batez ere, euskararekin gertatzen den bezala, ikasteko itzulpen adibide gutxiago dituzten hizkuntzetan.

Baina, esan bezala, azken urteetan itzulpena egiteko hurbilpen berria agertu da: Itzulpen automatiko neuronala. Aurreko sistema estatistikoak bezala, itzultzaile hauek ere, lehendik egindako itzulpen bildumetatik ikasten dute behar duten informazio guztia. Baina, aurreko sistemak ez bezala, hori egiteko neurona-sareak erabiltzen dituzte.

Neurona-sareak, izena eta inspirazioa giza neuronetatik eta garunaren egituratik hartu arren, metodo guztiz matematikoak dira. Labur esanda, metodo berri hauek hitzen esanahia bektoreen bitartez errepresentatzen dute eta egitura sintaktikoa matrizeen bidez. Esaldi asko eta bakoitzaren itzulpen zuzena dira abiapuntua. Itzultzen ikasteko prozesu automatikoan, hasieran matrize horiek ausazko datuekin hasieratzen dira eta esaldiak itzultzen dira automatikoki. Itzulpena nolakoa izan beharko lukeen dakigunez, hasieran ausazko balioak zituzten matrize horiek automatikoki egokituko ditugu, egindako errorea txikitzeko asmoz. Prozesu hau milioika esaldirekin errepikatu ondoren, itzultzailearen oinarri diren matrize horiek beste edozein esaldi itzultzeko prest egongo dira.

Itzultzaile automatiko neuronalek denbora gutxian garapen izugarria izan dute, eta egun Google, Microsoft eta Systran enpresek teknologia hau darabilte itzulpenak sortzeko, aurreko 20 urtetan garatutako teknologia estatistikoa alde batera utziaz. UPV/EHUko Ixa taldeak, MODELA (https://modela.eus/eu/itzultzailea)proiektuko beste kideekin lankidetzan (Vicomtech, Ametzagaina, Elhuyar eta ISEA) teknologia berri honetaz baliatzen den gaztelera-euskara itzultzaile bat sortu eta demo modura erabilgarri jarri dugu. Egindako ebaluazioetan itzultzaile berria orain arte zeunden guztiak baino hobea dela ikusi da, gaztelera eta euskararen arteko itzulpen automatikoa beste maila bat gorago eramanez. Deepl eta Google itzultzaileek hizkuntza nagusi gutxi batzuekin lortu duten maila euskararekin ere lortu dugu guk. Benetan berri pozgarria da.

Kalitatean hobekuntza nabaria lortu arren, sistema hauek sortzen dituzten itzulpenak ez dira beti zuzenak. Izan ere, sortzen dituzten esaldiak aurretik sortutako sistemenak baino askoz naturalagoak dira, baina ez dute beti jatorrizko esaldiaren esanahi bera mantentzen. Errore mota berri hau itzulpen automatikoaren erabiltzaileentzako erronka bat da. Aurreko sistemetan, kalitate txarreko itzulpenak berehala antzematen ziren, kasu horietan sortutako esaldia naturala ez zelako. Oraingo sistemek, berriz, arreta berezia eskatzen dute, lehenengo begiradan egokia dirudien esaldi batek ezegokia izan baitaiteke jatorrizko testuaren informazio bera mantentzen ez duelako.

MODELA sistemarekin itzulpengintza automatikoan lortu dugun hobekuntza kualitatibo hau HiTZ zentro berrian kokatuko dugu. IXA taldea 2018an Aholab ikerketa taldearekin elkartu da HiTZ Hizkuntza-Teknologiako Zentroa sortzeko eta martxan jartzen ari gara orain. Nazioarteko erreferentzia-zentro bat eraikitzen ari garen honetan itzulpen automatikoarekin batera beste lau lerro estrategiko hauek ere landuko ditugu: medikuntzako testu eta testu juridikoen tratamendu automatikoa, irakaskuntzarako ariketak prestatzeko laguntzak, eta sare sozialetako edukien azterketa. “Hodeiko” konputazioan eta “big data” mailan erabil ahal izateko moldatu eta zabalduko ditugu gure baliabide eta tresna konputazionalak. Adimen artifizialean beharrezko den superkonputagailuen erabilera masiboari atea zabalduko diogu.

Madrileko Industria, Energia eta Turismoko Ministerioak sustatzen duen ‘Plan de impulso de las Tecnologías del lenguaje (http://www.agendadigital.gob.es/tecnologias-lenguaje/PTL/Paginas/plan-impulso-tecnologias-lenguaje.aspx)’ planaren ofizina tekniko baten papera jokatuko du HiTZ zentroak (90 milioi euro mugitzen ditu plan horrek). Horrela bidea zabalduko dugu planaren barruko kontratu-lizitazioetan parte hartzeari. Esan bezala, itzulpen automatikoan lortu ditugun emaitza bikain hauek nazioarteko erreferentzia garela erakusten dute, eta urrats esanguratsu berriak lortzeko animatzen gaituzte.