Mikel Artetxe Zurutuza
Hizkuntzaren prozesamenduan
doktoregaia EHUn

Hizkuntzen mapa birtualak

Dozenaka mila koordenatu, ehunka dimentsio, ipar ala hegorik ez, eta hitzak hirien partez. Baldarrak eta xelebreak beharbada, baina mapak dira hauek ere, eta makinek erabiltzen dituzte, hizkuntzaren ozeanoan nabigatzeko. Eurei esker, ordenagailu batek bere kabuz ondoriozta dezake erreginak emakumezkoak direla edo *txakur* frantsesez *chien* esaten dela. Ongi etorri *embedding*en mundu zoragarrira. Bidaia hastera doa.

Bitxia da hizkuntza. Nola idazkera hala ahoskera aldetik, *zerri* eta *zorri* hitzak oso dira antzekoak, eta *zerri* eta *basurde* hitzak, berriz, zeharo desberdinak. Animalien artean, baina, zerriak askoz hurbilago daude basurdeetatik zorrietatik baino. Azken finean, hizkuntza bateko hitzen eta haien esanahien arteko lotura arbitrarioa da, eta zerri, zorri eta basurde kontzeptuen arteko erlazioak gure buruan baino ez dira bizi.

Googlen bilaketa bat egitean edo Sirirekin hitz egitean, ordea, guk hain barneratuta daukagun hori zailtasun handi bihurtzen da makinentzat. Eta, mendian edo errepidean galtzean egiten dugun bezalaxe, makinek ere mapak erabiltzen dituzte hizkuntzen labirintoan aurrera egiteko. *Embedding* esaten zaie mapa horiei, eta, haien bidez, pentsaezinak ziruditen lurretan barneratu da hizkuntzaren prozesamendua. Goazen gu ere, pausoz pauso, bidaia hori egitera.

Hirien mapetatik hitzen mapetara

Hizkuntzaren lurraldean barneratu aurretik, geldi gaitezen une batez gure mapa arruntei erreparatzeko. Funtsean, ezagutzen ditugun mapak lurralde baten errepresentazio grafikoak dira, kasuan kasu hiri bakoitzari puntu bat esleitzen diotenak, adibidez. Mapek zentzua izan dezaten, noski, puntu horiek ez daude edozein modutara sakabanatuta, baizik eta errealitatean ditugun distantziak errespetatzen ditu mapako kokapenak. Horrela, mapa batean, Paris hurbilago agertuko zaigu Bruselatik Moskutik baino, errealitatean ere Frantziako hiriburua gertuago baitago Belgikakotik Errusiakotik baino.

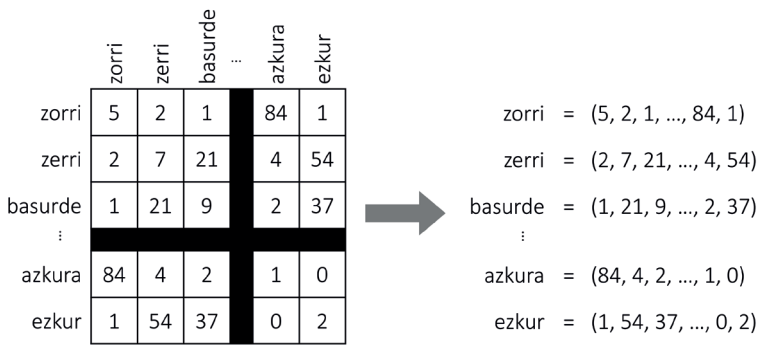
Hizpide ditugun hizkuntzen mapak ez dira oso ezberdinak. Hiriak azaldu beharrean, puntu bakoitzak hitz bat adierazten du, eta haien arteko distantziak hitzen arteko antzekotasun semantikoaren araberakoak dira. Horrenbestez, halako mapa batean, *zerri* hitzari dagokion puntua hurbilago egongo da *basurde*ri dagokionetik *zorri*ri dagokionetik baino, *zerri* eta *basurde* hitzen arteko antzekotasun semantikoa handiagoa baita *zerri* eta *zorri* hitzen artekoa baino.

Dena ez da hain erraza ordea: hizkuntzaren konplexutasuna behar bezala harrapatzeko, motz geratzen dira paperaren 2 dimentsioak, eta mapa hauek 300 dimentsio inguru izan ohi dituzte normalean. Baina ez zaitzatela beldurtu zenbaki han-

1. irudia. Euskarazko *embedding* batzuen bistaratzea tSNE teknika erabiliz.

diek! Lerro zuzenaren dimentsio bakarretik karratuaren bi dimentsioetara jauzi bat dagoen bezala, eta karratuaren bi dimentsioetatik kuboaren hiru dimentsioetara beste jauzi bat, imajina dezakezu badagoela antzeko jauzi bat kuboaren hiru dimentsioetatik teseraktoaren lau dimentsioetara, eta horrela jarrai genezake aipatutako 300 dimentsioetara iritsi arte.

Baina nola eraiki 300 dimentsioko mapa bat hiru dimentsiotan bizi bagara? Lasai, ez gara sorginkeriatan hasiko eta! Errealitatean, mapa horiek ez baitira fisikoak, ordenagailuen memorian bizi diren objektu matematikoak baizik. Izan ere, mapa guztiak zenbakien bidez errepresenta daitezke. Horretarako, erreferentzia-sistema bat adostu ohi da, eta puntu bakoitza ardatz ezberdinekiko duen posizioaren arabera adierazi. Hala, ekuatoreareki-ko eta Greenwich meridianoarekiko distantzia angeluarraren arabera, Parisen koordenatuak (48.86, 2.35) direla diogu, Bruselarenak (50.85, 4.35) eta Moskurenak (55.75, 37.62). Koordenatu horiek erabiliz, aukera dugu, besteak beste, hirien arteko distantziak matematikoki kalkulatzeko. Hizkuntzen mapekin ere antzera egiten da, baina, 2 dimentsio beharrean 300 dituztenez, 300 zenbaki behar dira puntu bakoitza deskribatzeko. Bada, hitz bat errepresentatzen duen zenbaki-segida horietako bakoitza da, hain justu ere, *embedding* esaten duguna.



2. irudia. *Embedding*en ikasketa, zenbaketa bidez. Testu luze bat hartuta, hitz-konbinazio bakoitza batera zenbat esalditan agertzen den kontatzen da taula batean. Errenkada bakoitzeko zenbaki-zerrendak dagokion hitzaren koordenatuak ematen ditu. Praktikan, hitzen maiztasuna egokitzeko eta dimentsionalitatea murrizteko teknikekin findu ohi da oinarrizko hurbilpen hori.

Testua oinarri, eta makinak kartografo

Mapak egitea lan neketsua da, inolaz ere. Kartografoek askotariko argazki, neurketa eta estatistikak bildu eta aztertu ohi dituzte, eta datu horiekin bat datozen errepresentazio grafikoak taxutu. Hizkuntzak deskribatzeko ere gizakiak egin izan ditu antzeko saiakerak: hortxe ditugu, besteak beste, hain arruntak zaizkigun hiztegiak. Baina hizpide ditugun mapak ez dira eskuz eginak. Testu luzeak aztertuz makinek eurek sortzen dituzte automatikoki, eta errezeta sinple bezain eraginkorra da.

2. irudian erakutsi bezala, demagun taula erraldoi bat eraikitzen dugula hizkuntza bateko hitz guztiekin. Hitz bakoitzarentzat errenkada eta zutabe bana izango dugu, eta, hala, gelaxka bakoitza hitz-bikote bati egokituko zaio. Taula betetzeko, testu luze bat hartuko dugu, eta hitz-bikote bakoitza batera zenbat esalditan ageri den kontatuko. *Et voilà*, hortxe dugu gure mapa! Hitz bakoitzeko, dagokion errenkadako zenbaki-zerrenda hartuko dugu, eta horiexek izango dira hitzaren koordenatuak.

Gezurra badirudi ere, hurbilpen sinple horrek nahiko mapa zentzudunak sortzen ditu. Izan ere, semantikako hipotesi distribuzionalaren arabera [8, 7], antzeko hitzek antzeko agerkidetza-patroiak izan ohi dituzte, eta, horrenbestez, aurreko prozedurak antzeko koordenatuak esleituko dizkie. Gure hasierako adibiderako ere, *basurde* nahiz *zerri* hitzak sarritan agertuko dira *ezkur* hitzarekin batera, eta gutxitan *azkura* hitzaren alboan; eta *zorri* hitzarekin alderantziz gertatuko da. Horrenbestez, *zerri* eta *basurde*k antzeko koordenatuak izango dituzte, eta elkarrengandik hurbil geratuko dira mapan. *Zorri*ren koordenatuak, berriz, nahiko ezberdinak izango dira, eta, ondorioz, haietatik urrunago egongo da.

Baina bada zerbait errezeta honetan falta dena. Izan ere, arestian aipaturiko 300 dimentsioek asko baziruditen ere, prozedura honek dozenaka mila dimentsioko mapak sortuko lituzke, hizkuntzak hitz adina zenbakiz osatuko baitira bertako koordenatuak. Nola murriztu, bada, dimentsio-kopurua? Erantzuna ez zaigu oso arrotza: gure mapa arruntek ere bi dimentsio izan ohi dituzte, nahiz eta errealitatean hiru dimentsioko mundu bat adierazi. Hain zuzen ere, mapak sortzean, kanpoan utzi ohi da altueraren dimentsioa, ez baita batere esanguratsua hirien arteko distantziak kalkulatzerako garaian. Hizkuntzen mapekin ere antzeko zerbait egiten da: hainbat teknika matematiko erabiliz, aldakortasun handieneko ardatzak (esanguratsuenak direnak) identifikatzen dira, eta gainerako dimentsioak mapatik kanpo uzten. Maiztasunaren efektua zuzentzeko moldaketa batzuk gorabehera, horixe da, hain justu ere, *embedding*ak ikasteko zenbaketa-tekni-

ken atzean dagoen oinarrizko ideia [3]; kontuan izan ikasketa automatikoan oinarritutako teknikek [11, 4] modu inplizituan egiten dutela prozedura bera [10].

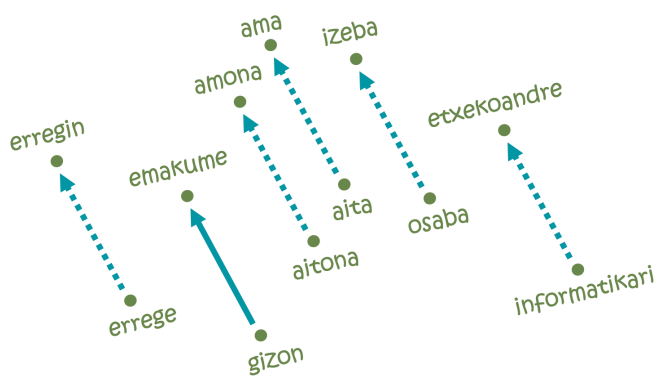
Mapekin jolasean

Hain sinpleak izanagatik, hasiera batean dirudien baino sekretu gehiago ezkutatzen dituzte gure mapa arruntek. Horretarako pentsatu izan ez badira ere, munduko edozein txokotan zer tenperatura egin ohi duen estimatzeko ere balio dute, adibidez. Izan ere, muturreko latitudeetan aurkitzen diren puntuak, poloetatik hurbilen daudenak alegia, hotzagoak izan ohi dira, eta ekuatoretik gertuago daudenak, berriz, beroagoak. Bada, latitudearen ardatza tenperaturarekin lotzen den modura, hizkuntzen mapetan ere antzeko ardatzak identifika daitezke hitzen polaritatea (positibotasun- eta negatibotasun-maila) eta bestelako ezaugarri batzuekin lotzen direnak [12]. Haiei esker, indar handia hartu dute azken aldian iritzien azterketa automatikoa egiteko aplikazioek.

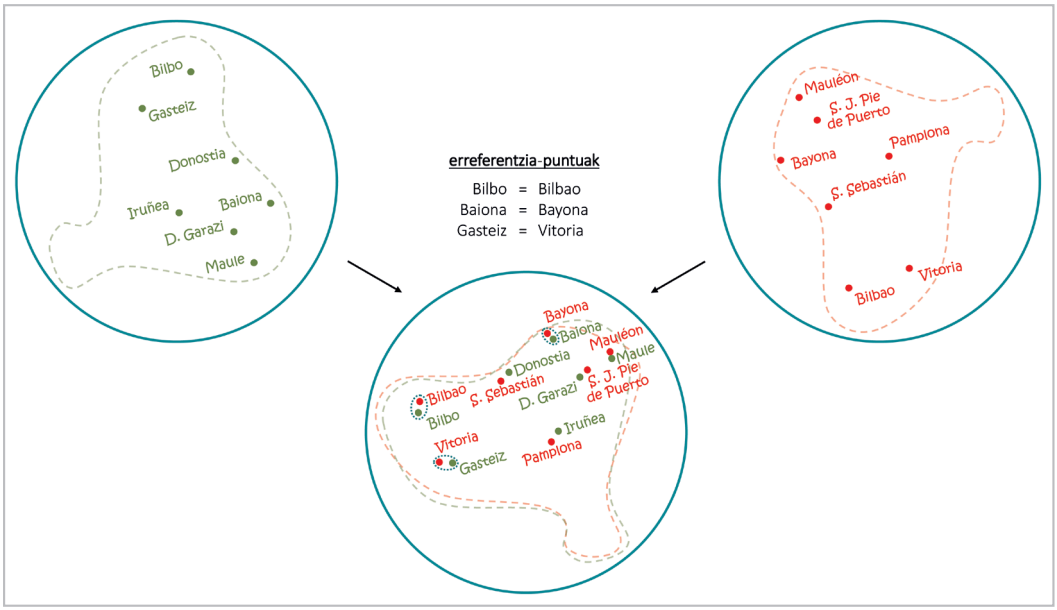
Baina, *embedding*ak hain ezagun egin dituena analogien ebazpena izan da [11]. Ideia ezin liteke sinpleagoa izan: Paristik Bruselara joateko, 222 km egin behar dira iparraldera eta 144 km ekialdera; era berean, hizkuntzaren mapako ardatz bakoitzean ere distantzia jakin bat beharko da *gizon* hitzetik hasita *emakume* hitzera iristeko, adibidez. Bada, *errege* hitzetik hasi eta pauso berberak ematen baditugu, *erregin* hitzera iritsiko gara! Izan ere, ikasitako ibilbideak gizon-emakume erlazioa kodetzen du, eta edozein hitz maskulinotatik abiatuta haren baliokide femeninora eramaten du horrenbestez. Horren antzera, analogia baliokideak egin daitezke herrialde-hiriburu, singular-plural, orainaldi-lehenaldi eta halako erlazioentzat.

Dena ez da hain polita, ordea: *informatikari* hitzetik hasita *gizon-emakume* ibilbide bera jarraituko bagenu, adibidez, *etxekoandre* hitzera iritsiko ginateke [5]. Bestela esanda, maparen arabera, informatika gizonen kontua da, eta etxeko lanak emakumeenak dira. Zer ikusi, hura ikasi: *embedding*ak gizakiek idatzitako testuetan oinarritzen direnez, gure gizartean errotutako joera diskriminatzaile berberak islatzen dituzte. Hain zuzen ere, hainbat adituren arabera, halako jokabide bidegabeei aurre egitea izango da adimen artifizialaren etorkizuneko erronketako bat.

Arazoak arazo, hizkuntza bakarrarekin halako trikimailuak egin badaitezke, hainbat hizkuntzatako mapak uztartuz are gauza harrigarriagoak lortu dira. 4. irudian ageri denez, euskarazko eta gazte-



3. irudia. Analogien ebazpena. Hitz maskulino batetik abiatu, eta, gizon-emakume ibilbidea eginda, haren baliokide femeninora iristen da.



4. irudia. *Embedding*en mapaketa. Ezberdin orientatutako Euskal Herriko bi mapa —bata euskarazkoa eta bestea erdarazkoa— elkarren gainean jartzen dira erreferentziazko hiri-bikoteak (Bilbo-Bilbao, Baiona-Bayona, Gasteiz-Vitoria) elkarrengana hurbilduz. Behin hori eginda, itzulpen-bikote berriak erauz daitezke gainjarritako mapan elkarren ondoan geratu diren puntuei erreparatuz (adibidez, Iruñearen ordaina Pamplona dela ondoriozta daiteke, dagozkien puntuak elkarren ondoan geratu direla ikusita). Hizkuntzen mapekin ere oinarrizko printzipio hori baliatzen da hitz arrunten itzulpenak induzitzeko.

laniazko mapa bana gainjarriz Euskal Herriko hiriburuen erdal ordainak erauz daitezkeen bezala, *embedding*ekin ere oinarrizko printzipio bera baliatzen da hitz arrunten itzulpenak induzitzeko [1, 6]. Bide horretatik, gizakien inolako gidaritzarik gabe ikasteko gai diren itzultzaile automatikoak garatu dira berriki [2, 9], hainbat hizkuntzatan idatzitako testu luzeak irakurri eta, bestelako laguntzarik gabe, haien arteko itzulpenak egiteko gai direnak.

Helmuga berriak

Gure bidaia bukaerara heltzear bada ere, *embedding*ek zabaldutako bideak amaigabea dirudi. Ikasketa-teknikak hobetu eta aplikazio berriak garatzearekin batera, hitzen mapetan oinarrituz esaldi edo testu luzeagoei heltzeko saiakerek hartu dute indarra azkenaldian. Bide horrek noraino eramango gaituen, baina, ez da inongo mapetan ageri, eta, helmuga berriak izanik zeruertzean, etorkizuna ezin zitekeen zirraragarriagoa izan. ●

Erreferentziak

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017.

[2] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In Proceedings of the Sixth International Conference on Learning Representations, 2018.

[3] Marco Baroni, Georgiana Dinu, and German Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146, 2017.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in Neural Information Processing Systems 29, 2016.

[6] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herve Jegou. Word translation without parallel data. In Proceedings of the Sixth International Conference on Learning Representations, 2018.

[7] J.R. Firth. A Synopsis of Linguistic Theory, 1930-1955. 1957.

[8] Zellig S Harris. Distributional structure. Word, 10(2-3):146{162, 1954.

[9] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In Proceedings of the Sixth International Conference on Learning Representations, 2018.

[10] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In Advances in Neural Information Processing Systems 27, 2014.

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Je_ Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26, 2013.

[12] Sascha Rothe and Hinrich Schütze. Word embedding calculus in meaningful ultradense subspaces. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016.

CAF-Elhuyar sarietara aurkeztutako lana.