Gorka Azkune Galparsoro
Ikertzailea eta irakaslea Deustuko
Unibertsitateko Ingeniaritza Fakultatean

Zenbakien hizkuntza

Zientziak eta letrak. Zenbakiak eta hitzak. Bi mundu bereizi balira bezala aurkeztu izan zaizkigu ia beti. Elkarrekin dantza egin nahi ez duten izaki isolatu gisa. Hala eta guztiz ere, gure garunek badakite bai zenbakiak bai hitzak behar bezala lotzen. Zenbakiak eta hitzak darabiltzagu unibertsoa ulertzeko, haren misterioak bizitzeko, gizakion sentimendu eta nahiak adierazteko. Agian, hor, haien sorreran, topatu beharko genuke haien arteko lotura: adimenean.

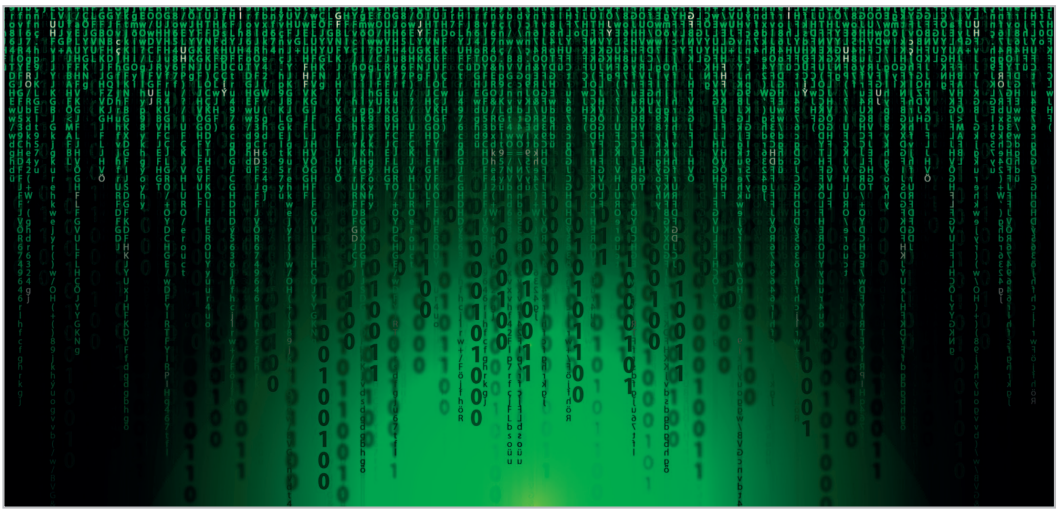
Giza adimena oso konplexua da. Hain konplexua, ezen oraindik ez dugun haren azala urratu besterik egin. Nolanahi ere, argi dago adimen horren oinarrietako bat hizkuntza dela. Hizkuntzak ahalbidetu digu kontzeptu konplexuak adieraztea, ideiei forma eman eta gure kideei transmititzea, kultura aberatsak egituratu eta hurrengo belaunaldietan aztarna uztea.

Hizkuntzak gure adimenean zenbaterainoko garrantzia duen kontuan hartuta, funtsezko ikergaia bilakatu da adimen artifizialaren esparruan. Lengoaia naturalen prozesamendua (LNP) deitzen diogu, eta askok uste baino aplikazio gehiagotan ikus dezakegu: LNPa erabiltzen da itzultzaile automatikoetan, *spam* mezuak identifikatzeko eta Amazonen erosi dugun produktu baten iruzkina sailkatzeko.

Azken urteetan, LNParen esparrura ere iritsi da sare neuronal artifizialen iraultza, eta ondorio praktiko ikusgarriak eragin ditu [1]. Baina ez hori bakarrik: hitzak eta zenbakiak elkarrekin nola lotzen diren jakiteko aztarna interesgarriak azaleratu dizkigu.

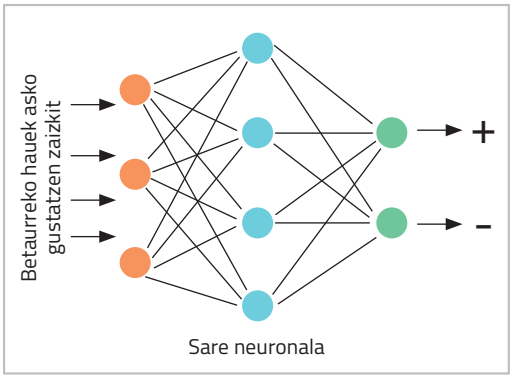
Makinentzat hitzak irudikatzen

Jo dezagun betaurrekoak egiten dituen enpresa bateko marketin-zuzendariak garela. Betaurreko berri batzuk merkaturatu ditugu, eta erosleen iritziak



jaso nahian gabiltza. Horretarako, Twitter sare soziala erabiltzea otu zaigu. Gure betaurreko berrien izena daraman traola sortuko dugu, jakiteko gure bezeroek betaurrekoei buruz zer idazten duten. Arazotxo bat dugu, ordea: gure enpresak mundu guztian saltzen duenez, txio asko espero ditugu. Beraz, ezin txio guztiak irakurri! Lan hori guztia ordenagailuak egitea nahiko genuke.

Zehatz ditzagun lanak hobeto: txio bat hartuta, gure ordenagailuak erabaki behar du txioan gure betaurrekoei buruz iritzi positiboa ala negatiboa ematen den. Has gaitezen lanean bada. Halako arazoak egokiak dira sare neuronalentzat. Ideia sinplea da: erakutsi sare neuronalari hainbat txio, eta adierazi ea positiboak ala negatiboak diren; adibideak ikustearen poderioz, sareak ikasiko du txio positiboak eta negatiboak bereizten (1. irudia).



1. irudia. Sare neuronal bat nola erabili, txio baten iritzia sailkatzeko.

Baina badugu beste arazotxo bat: sare neuronalek zenbakiekin egiten dute lan, ez hitzekin. Nola irudikatu behar ditugu hitzak zenbakiak erabiliz? LNParen munduan, lan handia egin da horren inguruan, eta, horren ondorioz, hitzak irudikatzeko modu asko daude. Ideia sinpleena da ingelesez *one-hot vector* deritzona.

Demagun lau hitz irudikatu nahi ditugula: *goi, behe, ezker* eta *eskuin*. Horretarako, lau zenbakiko bektoreak erabiliko ditugu, eta hitz bakoitzari posizio bat esleituko. Gure lau elementuko bektorea zeroz osatuta egongo da, hitz bakoitzari esleitu diogun posizioan izan ezik; hor bateko bat jarriko dugu (2. irudia).

Goi	→	[0, 0, 0, 1]
Behe	→	[0, 0, 1, 0]
Ezker	→	[0, 1, 0, 0]
Eskuin	→	[1, 0, 0, 0]

2. irudia. *One-hot vector* irudikapenaren adibidea.

Irudikapen-mota hori oso sinplea da, eta ongi bereizten du hitz bakoitza, baina baditu hainbat desabantaila. Adibidez, denok dakigu *goi* eta *behe* hitzak antonimoak direla. Bi hitz horien bektoreek erakusten al dute erlazio hori? Ez. *One-hot vector* delakoak erabiliz gero, ezin dira irudikatu hitzen arteko erlazioak. Beste arazo bat: euskarak, esaterako, 37 mila hitz inguru ditu [2]. Horiek denak irudikatzeko 37 mila dimentsioko bektoreak beharko genituzke! Ez dirudi, beraz, oso ideia ona denik.

Word2Vec

Ideia erakargarria da bektoreak erabiltzea hitzak irudikatzeko, sare neuronalek ongi egiten baitute lan bektoreekin. Baina *one-hot* bektoreak hobetu beharrean gaude. Hori bera pentsatu zuten Mikolovek eta haren lankideek *Word2Vec* teknika asmatu zutenean [3]. Teknika horren bitartez, hitzen irudikapen oso interesgarriak lortzen dira, honako arrazoi hauengatik:

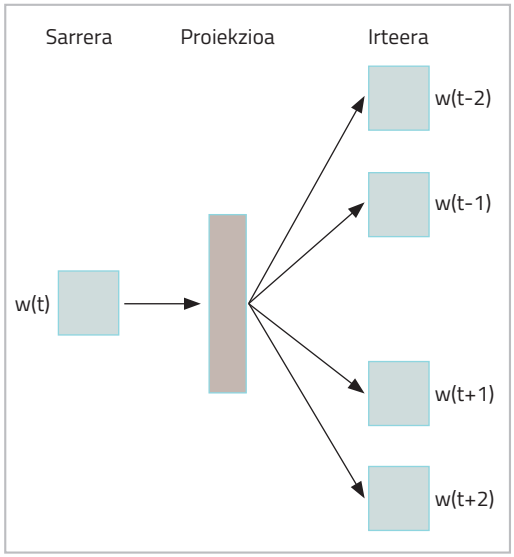
1. Hitzak bektore txikien bidez irudika daitezke.
2. Hitzen arteko erlazio semantikoak irudika daitezke.

Nola lortzen da, ordea, hori egitea? Sare neuronalak modu bitxi batean erabiliz. Demagun euskarazko Wikipedia osoa hartzen dugula testu gisa. Bertan,

euskarazko hitz gehienak agertuko dira, ongi egituratutako esaldietan. Esaldi horietako hitzak hartuko ditugu, eta *one-hot vector* irudikapen sinplea erabiliz kodetuko ditugu. Orain dator amarrua. Hitz bat hartuko dugu, sare neuronal bati pasatzeko, haren helburua izanik hitz horren aurretik eta atzetik dauden bi hitzak asmatzea. Hau da, hitz bat emanda haren testuingurua asmatzeko entrenatuko dugu sare neuronala, 3. irudian dakusagun moduan.

4. irudian ikus dezakegu entrenamendu hori egiteko zer sare neuronal erabiltzen den. Gure euskarazko Wikipediako testu osoa prozesatuko du, erakusten den moduan. Hitz bat emanda, haren testuingurua asmatzeko gaitasuna landuko du entrenamendu horretan. Hasieran, testuinguru-hitzak ez ditu asmatuko; beraz, errore handiak egingo ditu. Baina errore horiek sarea entrenatzeko erabiltzen dira. Horrela, milioika hitzez osaturiko testuak ikustearen poderioz, sareak bere asmatze-erroreak gutxituko ditu.

Prozesu horretan zehar, ordea, non daude ikasitako hitzen irudikapenak? 4. irudian *proiekzioa* izenpean irudikatu dugun laukitxoan, hain zuzen ere. Agian, hobeto ulertzeko, 5. irudiari begiratu beharko diogu. Bertan ikus daiteke sare neuronalaren egitura zehatzagoa, bi hitz bakarrik hartu diren kasu batean: sarrerako bat eta irteerako beste bat. Entrenamendu-prozesua amaitzen denean, hitz baten irudikapena lortzeko, nahikoa da entrenatu berri dugun sare neuronalari hitz hori pasatzea eta geruza ezkutuko aktibazioak hartzea. Gure garunetako neuronak bezalatsu, neurona artifizialak ere modu



4. irudia. Sare neuronal batentzako entrenamendu-prozesua, hitzak nola irudikatu ikasteko. Hitz bat sarrera gisa emanda, haren aurretik eta ondoren dauden bi hitzak asmatzen ikasten du.

ezberdinetan aktibatzen dira estimulu ezberdinen arabera. Bada, hitz ezberdin batentzat lortzen diren aktibazioak izango dira hitz horren irudikapen egokia. Ez al da harrigarria?

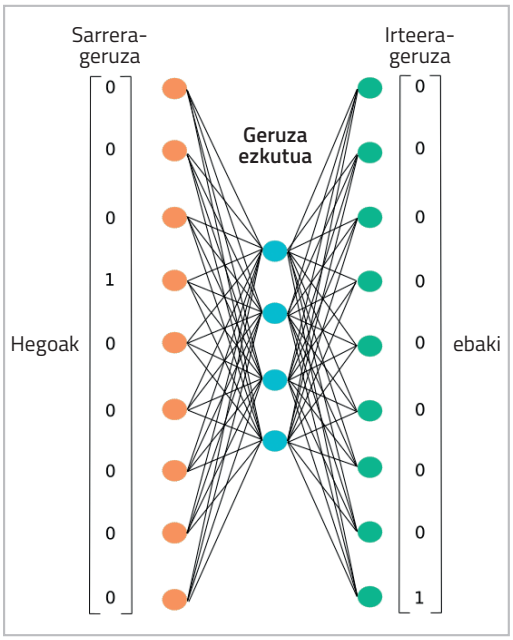
Azken batean, ataza jakin bat egiteko entrenatu dugun sare neuronalak, modu automatikoan, hitzen zenbakizko irudikapen batzuk ikasi ditu. Eta irudikapen horiek oso ahaltsuak dira.

Hitzekin jolasean

Geruza ezkutuko neurona horien aktibazioak zenbakiak dira. Beraz, geruza horretan 300 neurona

Esaldia	Sarrera-hitza	Testuingurua
Hegoak ebaki banizkio nerea izango zen	Hegoak	__, __, ebaki, banizkio
	ebaki	__, Hegoak, banizkio, nerea
	banizkio	Hegoak, ebaki, nerea, izango

3. irudia. Esaldi bat hartuta esaldiko sarrera-hitz batzuentzat sortzen diren testuinguruak.



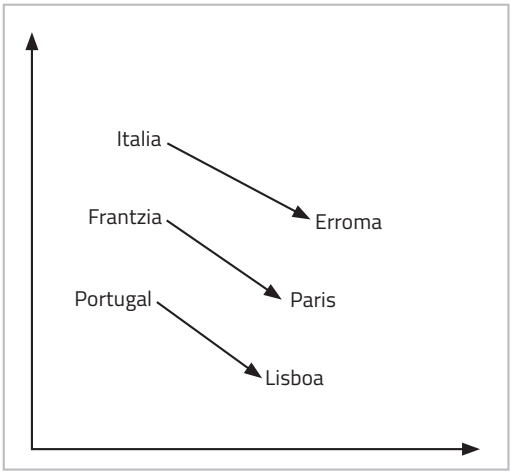
5. irudia. *Word2Vec* sare neuronalaren irudi zehatzagoa. Geruza ezkutuko neuronen aktibazioek osatzen dute "hegoak" hitzaren irudikapena.

jartzen baditugu, 300 zenbakiz osaturiko bektore bat lortuko dugu edozein hitzentzat. *One-hot vector* irudikapenean 37 mila zenbakitik gora behar genituela kontuan hartuz, asko irabazi dugu, ezta? Baina hori ez da onena. Hitz-bektore berri horiek propietate ia magikoak dituzte. Jolas gaitezen haiekin; aho-zabalik geratuko zarete!

Has gaitezen galdera batekin: Italiarentzat Erroma dena, zer da Frantziarentzat? Zure erantzuna Paris izango da. Zergatik? Italia-Erroma erlazioan herrialdea eta hiriburua ikusi dituzulako. Haien esanahia erabiliz, arrazoiketa hori egin duzu. Beraz, Frantziaren hiriburua zein den pentsatu duzu: Paris. Galderaren erantzuna topatzeko, beharrezkoak izan dira lengoaia eta kontzeptuak maneiatzea. Prozesu nahiko konplexua dirudi.

Ikasi berri ditugun hitz-bektoreek gaitasuna ematen digute era horretako galderei erraz erantzuteko. Kasu honetan, nahikoa da Italia – Erroma + Frantzia eragiketa egitea. Hots, Italia bektoreari Erroma bektorea kendu, eta, ondoren, Frantzia bektorea gehituko diogu. Eta bai, emaitza Paris bektorea da! Beste adibide bat: Errege – gizon + emakume = erregina. Txundigarria, ezta?

Nola da hori posible? Bektoreak gehitu eta kentzea arrazoitzearen parekoa ote da, bada? Ikasi ditugun hitzen irudikapenak 300 dimentsioko espazio bateko puntuak besterik ez dira. Pentsa dezagun bi dimentsiotan, errazago ikusteko. 6. irudian herrialde eta hiriburu batzuk irudikatu ditugu 2 dimentsioko plano batean. Ikus daitekeenez, herrialdeak elkarrengandik gertu agertzen dira, baita hiriburuak ere, halako kategoria semantiko baten kide baitira. Baina, gainera, herrialde baten eta haren hiriburuaren arteko distantzia berdintsua da herrialde-hiriburu pare guztientzat. Horregatik funtzionatzen dute gure batuketek eta kenketek. Hori bera pasatzen da hizkuntza bateko hitz guztiekin, baina 300 dimentsioko espazio batean. Azkenean, hitzen arteko



6. irudia. Bi dimentsioko espazioan, herrialdeak eta hiriburuak nola antolatzen diren.

erlazioak, esanahia eta ñabardurak propietate geometrikoak baino ez dira.

Kontu egin inork ez duela diseinatu hitzen irudikapen horiek osatzen duten espazio semantiko hori. Sare neuronal batek ikasi du, bere kasa; hitz bat emanda haren testuingurua asmatzeko entrenatu den sare neuronal batek. Harrigarria da pentsatzea nola ikasten dituen halako irudikapen konplexuak ataza sinple bat ikasteko entrenatzen dugun sare batek. Baina halaxe gertatzen da.

Amaitzeko

Sare neuronalen bidezko hitzen irudikapena da gaur egungo LNParen oinarria. Adibidez, itzulpen automatikoan erabiltzen diren sare konplexuek sarrera gisa hartzen dituzte hitz-bektore horiek. Euskarazko bektoreak hartu eta ingelesezkoak sortu, esaterako. Prozesu horietan ikusi da oso antzekoak direla hizkuntza ezberdinetan ikasitako hitzen arteko propietate geometrikoak. Hau da, ingelesezko *king, man, woman* eta *queen* bektoreen posizio erlatiboak euskarazko *errege, gizon, emakume* eta *erregina* bektoreen ia berdinak dira. Bide horiek aztertzen eta itzulpen-teknika berritzaileak proposatzen ari dira, besteak beste, EHUko Ixa Taldeko, Elhuyarreko eta Vicomtecheko hainbat ikerlari [4].

Beste ikerlari batzuk erakutsi dute gure hizkuntzan ditugun joera sexistak hauteman daitezkeela bektoreen propietate geometrikoetan. Horrela, hitz horien joera sexistak ezabatzeko teknikak proposatu dituzte, makinek gure hutsegiteak errepika ez ditzaten [5].

Lan horiek guztiek goitik behera aldatzen dute hitzen eta zenbakien arteko erlazioari buruz dugun ikuspegia ere. Kontu izan gure neuronek tentsio elektrikoarekin egiten dutela lan. Zenbakiekin, nolabait. Neuronen arteko elektroien joan-etorrietan daude gure oroitzapenak, sentimenduak, hizkuntza eta arrazoimena. Sare neuronal artifizialekin gertatzen den bezalatsu, testu hau idaztean ez ote naiz ari, konturatu gabe, zenbaki-pilo batekin eragiketak egiten? Ez ote da gurea zenbakien beste hizkuntza bat besterik? ●

CAF-Elhuyar sarietara aurkeztutako lana.

Bibliografia

[1] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553 (2015): 436-444.

[2] Euskararen hitz kopurua: https://31eskutik.com/2016/06/03/zenbat-hitz-ditu-euskarak/ (azken bisita: 2018/01/20).

[3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

[4] Artetxe, M., Labaka, G., Agirre, E., & Cho, K. (2017). Unsupervised neural machine translation. arXiv preprint arXiv:1710.11041.

[5] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in Neural Information Processing Systems (pp. 4349-4357).