IKER GAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

II. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2017ko maiatzaren 10, 11 eta 12
Iruñea, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Testu-loturen labirinto semantikoan barna, esanahi-bektoreak lagun!

Josu Goikoetxea, Iñigo Lopez-Gazpio, Eneko Agirre, Montse Maritxalar eta *Aitor Soroa*

75-82 or.
https://dx.doi.org/10.26876/ikergazte.ii.03.11

ANTOLATZAILEA: udako euskal unibertsitatea
ELKARLANEAN: upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
LAGUNTZAILEAK: eman ta zabal zazu UPV EHU · euskarabidea · MONDRAGON UNIBERTSITATEA · Nafarroako Gobernua Gobierno de Navarra · Ayuntamiento de Pamplona Iruñeko Udala · Deusto Universidad de Deusto Deustuko Unibertsitatea · EUSKAL HERRIA UDALBILTZA · Universidad de Navarra

Testu-loturen labirinto semantikoan barna, esanahi-bektoreak lagun!

Goikoetxea, Josu* eta **Lopez-Gazpio,Iñigo***
eta Agirre, Eneko eta Maritxalar, Montse eta Soroa, Aitor

IXA taldea. UPV/EHU.
{josu.goikoetxea,inigo.lopez,e.agirre,montse.maritxalar,a.soroa}@ehu.eus

Laburpena

Lan honetan Hizkuntzaren Prozesamenduaren alorreko bi ikerketa-lerro aurkezten ditugu: 1) semantika distribuzionala eta bektore-espazioen konbinaketa, eta, 2) testu-lotura eta honek irakaskuntzan duen erabilgarritasuna. Artikuluan zehar bi alor hauen artearen egoera kokatu, teknikak azaldu eta etorkizuneko erronkak planteatzen dira.

Hitz gakoak: Hizkuntzaren Prozesamendua, Semantika Distribuzionala, Bektore-espazioak, Testu-loturak

Abstract

In this paper we present two works carried out in the field of Natural Language Processing: 1) distributional semantics and vector space combinations, and, 2) textual entailment and its usefulness in the educational scenario. Throughout the article we describe the state of the art of these fields, analyze methods and discuss future challenges.

Keywords: Natural Language Processing, Distributional Semantics, Vector Spaces, Textual Entailment

1 Sarrera eta motibazioa

Lan hau *Hizkuntzaren Prozesamendu*aren (*HP*) esparruan kokatzen da, ingelesez *Natural Language Processing* gisa ezagutzen den arloan. Ikerketa-lerro zeharo zabala dugu *HP*a, hainbat diziplina konbinatzen baititu, bereziki: informatika, adimen-artifiziala eta hizkuntzalaritza. Adibide batzuk aipatzearren, hurrengo ataza hauek guztiak *HP*an ikertzen dira: hitzen analisi lexikoa, morfologikoa eta sintaktikoa, hitzen adiera-desanbiguazioa, itzulpen automatikoa, diskurtsoaren analisia, hitzen analisi semantikoa, hizkuntza-sorkuntza, etab. Oro har, *HP*a hizkuntzen eta makinen arteko interakzioarekin lotzen da, eta, ondorioz, *HP*ko adituen helburu nagusia hizkuntza ulertzeko eta hizkuntza sortzeko gai diren sistema adimendunak garatzea da.

Artikulu honetan hizkuntza ulertzeko metodo automatikoak aztertuko ditugu, hau da, makinek hizkuntzak interpretatzeko egiten duten prozesamendu automatikoaz arituko gara. Artikuluan ikusiko dugun moduan, makinei hizkuntza ulertzen edo interpretatzen irakatsi behar zaie, eta, helburu hau lortzeko, hitzak zenbakien bitartez adierazi beharko ditugu. Hala, hitzak zenbakien bitartez adierazita egongo dira (2.1 atalean azalduko diren metodoei esker), eta, era berean, zenbakiak makinak ulertzeko gai diren seinale elektrikoen bitartez (sistema bitarrari esker). Beraz, *HP* automatikoa egite aldera, makinek hitzen kontzeptuen errepresentazio abstraktuekin lan egin behar dute, eta azken hori ahalbidetzen duen metodoak *semantika distribuzionalean* (*Distributional Semantics*) oinarritzen dira. *Semantika distribuzionala* ia *HP*ko ikerketa-lerro guztietan erabiltzen da, eta gure lanaren abiapuntua izango da.

Bada, *semantika distribuzionala*ren muina *Hipotesi Distribuzionala* deiturikoa da (Harris, 1954); hots, *esanahi antzeko hitzek testuinguru berean agertzeko joera dute*. Ikus dezagun adibide bat, demagun *txadeos* hitzaren esanahia ez dakigula, hala ere, hurrengo esaldiak irakurrita gizakiontzat erraza da bere esanahia inferitzea:

[0]Ikergazteak diren autoreak * batez adierazita daude.

1. *Txadeos* edalontzia hartu zuen.
2. Peiok itzelezko mozkorra harrapatu zuen *txadeos* larregi edatearren.
3. Malbec, *txadeos* mahats ezezagunenetakoa, oso ondo egokitzen da Australiako eguraldi eguzkitsura.
4. Edariak gozo-gozoak ziren; *txadeos*a gorria eta gogorra, eta Rhenisha argia eta suabea.

1-4 esaldiak irakurrita, badakigu *txadeos*a mahatsarekin eginiko edari alkoholiko gogorra eta gorria dela. Aipatu berri dugun *Hipotesi Distribuzionala* oinarri harturik *Eredu Semantiko Distribuzional*ak (*ESD*) garatu dira *HP*an. 2.1 eta 3.1 ataletan azalduko dugunez *ESD*ak erabilita hitzen esanahiak modu abstraktuan errepresentatu daitezke espazio euklidear bateko puntu gisa. Era honetan, hitz bakoitza zenbakizko bektore baten bitartez islatuta egongo da, eta azken hori hitzaren *esanahi-bektore*a (*EB*) izango da. Adibide gisa, 1 irudian lau hitzen esanahiak bi dimentsiotako espazio batean irudikatuta daude.

Oro har, hitzak *EB*tan adierazita izateak hainbat abantaila ematen dizkigu, marko matematiko batean mugitzeko aukera ematen digulako. Honi esker, posible da adimen artifizialeko teknika desberdinak erabiltzea hitzen banakako esanahiak konbinatu eta esaldien errepresentazioak lortzeko (bektoreen gaineko eragiketak erabiliz). Era honetan, hitz mailako atazak ez ezik, esaldi mailako atazak ere landu daitezke, hala nola: esaldi pareen arteko antzekotasunak, desberdintasunak edota erlazioak identifikatzea.

Esaldi pareen lotura logikoen analisi automatikoari *Recognition of Textual Entailment* (*RTE*) edo *Natural Language Inference* (*NLI*) deritzo, eta oso ataza interesgarria da *HP*an hezkuntzarekin lotura handia duelako. Izan ere, *RTE* edo *NLI* zehaztasunez burutzeko gai den sistema bat ikasle bati feedback esanguratsua emateko gai litzateke, azterketa, galdera, ariketa, etxekolan, edo antzekoren baten aurrean. Hau dela eta, sistema interesgarriak dira hezkuntzan irakaslearen rola edo papera hartzera bideratuta baitaude, esaterako: internet bidez ematen diren online ikastaro masiboetan (*Massive Open Online Course* edo *MOOC*). 2.2 eta 3.2 ataletan sakonduko dugu testu-loturen metodoen inguruan.

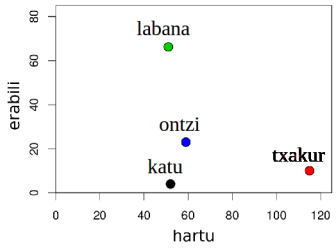
2 Arloko egoera eta ikerketaren helburuak

2.1 Esanahiaren errepresentazio distribuzionala HPan

*ESD*ak, modu inplizituan bada ere, $n \times m$ tamainako M ko-okurrentzia-matrize erraldoi batetik abiatzen dira; M, beraz, corpus bateko berben arteko ko-okurrentziez osatutako matrizea da, non m lerroak xede-hitzak diren, eta n zutabeak tasunak edo dimentsioak. 1. irudiko matrizeak, adibidez, testu-corpus batean lau izenek (*labana*, *katu*, *txakur* eta *ontzi*) bi aditzekin (*hartu*, *erabili*) izandako ko-okurrentziez osatuta dago.

1 Irudia: Ko-okurrentzia martrizea, eta bere irudikapena bi dimentsiotan

M	hartu	erabili
labana	**51**	**84**
katu	52	4
txakur	115	10
ontzi	**59**	**23**



1 irudiko hirugarren lerroan *txakur* hitzak *hartu* eta *erabili* aditzekin izandako ko-okurrentziekin x_{txakur} bektorea osatu dezakegu. Beste berba guztiekin ere, noski, gauza bera egin dezakegu. Ko-okurrentzia horiek, nolabait esateko, hitzek espazio euklidear batean dituzten koordenatuak dira, non, Hipotesi Distribuzionalari jarraiki, testuinguru antzekoak dituzten hitzak gertu egongo diren espazio horretan, eta zerikusirik ez duten hitzak, berriz, urruti. 1 irudian aurreko paragrafoan aipatutako lau hitzak *hartu* eta *erabili* dimentsioetan ilustratu dira. Artikulu honetan espazio horri bektore-espazio deituko diogu.

*ESD*etan, oro har, M ez da bere horretan erabiltzen, eskala-faktoreren bat aplikatzen zaie edo transformazioren bat. Hainbat atazatan erabili izan dira *HP*an; hala nola, TOEFL sinonimia atazan, tesauroen

sorkuntza automatikoan, Part Of Speech atazetan, edota garuneko neurona-aktibazioak aurresateko.

Azken urteetako konputazio-ahalmenaren igoeraren eskutik, hamarkadatan ahaztutako neurona-sareen erabilera hainbat diziplinatan bogan jarri da. Neurona-sareen berpizkunde horrekin, *ESD*ekin hertsiki lotutako neurona-sare hizkuntza ereduak (*NSHE*) arrakasta handia izaten ari dira, emaitza oso onak lortzen baitituzte arestian aipatutako atazetan. Artikulu honetan *word2vec* (Mikolov *et al.*, 2013) *NSHE*aren bektoreak erabiliko ditugu gure atazako sarrera gisa.

Bada, hizkuntza ereduek ele jakin bateko hitzen sekuentzien tasun estatistikoak jasotzen dituzte; aurreko berbak jakinik, hurrengoa aurresateko gai dira. Gauzak horrela, *NSHE*ak neurona-sareetan oinarritutako hizkuntza-ereduak dira, eta neuronen aktibazioak hitzen errepresentazio distribuzionalak (esanahiak) legez erabiltzen dituzte. Corpus bat prozesatu ostean, berba bakoitzari d dimentsiotako bektore bat esleitzen diote; hain zuzen, errepresentazio horiek dira EBak, balio eskalarrez beteriko bektore trinkoak. Aurreko paragrafoetan esandakoaren harira, EBen dimentsio bakoitza tasun semantiko edota sintaktiko abstraktua da, eta, euren bektore-espazioaren dimentsio kopurua *NSHE*aren neurona kopuruak definitzen du. EBen dimentsioa, oro har, 50 eta 500 artean dabil.

Kontuan izan ko-okurrentzia matrizeek $V \times V$ tamaina daukatela, V hiztegiaren hitz kopurua izanik. Gauzak horrela, berba baten esanahia (matrizeko lerro bat) definitzeko milaka edo milioika dimentsiotako bektorea da (hitz horrek corpuseko beste berba guztiekiko dituen ko-okurrentziak). *NSHE*etan, ordea, matrizeak $V \times d$ tamaina du, eta, hortaz, bektoreek informazioa modu askoz konpaktuagoan kodetzen dute. *txakur* hitzaren EBak, esaterako, hurrengo itxura izango luke *NSHE* batean:

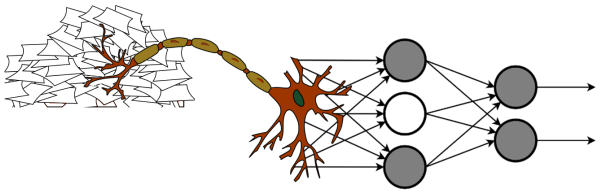
txakur 0.0214 0.2331 -0.5431 0.0022 -0.3411 ...

3.1 atalean aipatutako *word2vec*ek M matrizea ez du modu esplizituan erabiltzen, inplizituan inplizituan baizik.

2.2 Hezkuntzari lotutako HPa

Hezkuntzari lotutako HP (*Educational NLP*) betidanik interes handia piztu duen ikerketa-lerroa da. Ataza bere osotasunean ebaztea batere tribiala izan ez arren, aspalditik da komunitate zientifikoa arazo zehatz batzuei aurre egiteko metodoen eta estrategien bila. Esate baterako, 2000. urteaz geroztik adin desberdinetako ikasleek galdera laburrei emandako erantzunak automatikoki ebaluatzeko atazak pil-pilean egon dira. Leacock eta Chodorow; Nielsen *et al.* autoreek lan aipagarriak egin zituzten arlo honetan 2003; 2008 urteetan, hurrenez hurren. Esalditik testura salto egiten badugu, Zechner *et al.* autoreen 2015. urteko lana eta McNamara *et al.* autoreen 2015. urteko lana aipatzea ezinbestekoa da. Lan hauetan guztietan ikasleen idazlan laburrak interpretatzeko eta ebaluatzeko gai diren sistema automatikoak deskribatzen dira. Ikerketa hauen ondorio nagusia hurrengoa da: zenbat eta informazio gehiago txertatu sisteman orduan eta hobekiago interpretatu daitekeela testua. Hau dela eta, hasiera batean azaleko sintaxitik zein testutik zuzenean erauzitako ezaugarri sinpleak erabiltzen baziren ere; geroz eta arkitektura konplexuagoak erabiltzen hasiak dira ikertzaileak. Gaur egun, *Semantika distribuzionala*ren eta *EB*en arrakastarekin bat eginez neurona sareetan oinarritzen diren arkitektura konplexuak erabiltzen dira gehienbat (ikusi 2 irudia).

2 Irudia: Neurona-sareetan oinarritutako sistemek sarrera testua prozesatzen dute (*EB*ak erabiliz) eta atazaren araberako irteera bat ematen ikasten dute: esaldien arteko antzekotasunak eta desberdintasunak, kalifikazio bat, etab



Neurona-sareetan oinarritutako sistema konplexuen arazo nagusienetako bat da izugarrizko datu pila prozesatu behar dituztela beren emaitzak zehatzak izateko. Hori dela eta, ikasketarako datu-baseak ezinbesteko sarrera iturri bilakatu dira. Behar honi erantzuteko asmoz hainbat nazioarteko ataza sortu dira urteetan zehar, horien artean *Semeval workshop*ean 2013. urtean izandakoa. Bertan Dzikovska *et al.*

autoreek aurrerapauso handia eman zuten irakaskuntzaren domeinuko ataza bat komunitate zientifikoan zabaldu baitzuten. Ataza honek ikasleen eta irakasleen anotazioak biltzen zituen datu-base erraldoi bat atzigarri uzten zuen ikerketarako, eta ikerketa zentro askok erabaki zuten bertan parte hartzea beraien sistema propioak garatuz. Ataza honekin batera, eta lehen aldiz, autoreek *RTE/NLI* azpiataza bat ere proposatu zuten. Azpiataza honetan sistemak gai izan behar ziren erreferentziazko esaldiko kontzeptuak eta ikasleen erantzunetako kontzeptuak automatikoki lotzeko, eta loturen (ingelez *mapping* deritzona) araberako kalifikazio bat esleitzeko (ikusi 3 irudia).

3 Irudia: Semeval 2013. RTE/NLI atazako adibide bat (euskaratua). Bertan sistemek ikasleek emandako erantzunak kalifikatu behar dituzte erreferentzia esaldi batekiko.

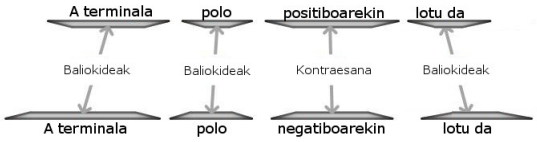
```
[Galdera:] Zergatik neurtu da 4,5 volteko tentsio-erorketa A eta B terminalen artean?

[Erreferentzia 1:] A terminala polo positiboarekin lotu delako.
[Erreferentzia 2:] B terminala polo negatiboarekin lotu delako.

[Ikaslearen erantzuna 1:] Ez dakit. (kalifikazioa: Domeinuz kanpo.)
[Ikaslearen erantzuna 2:] A eta B terminalak polo berriekin lotu direlako. (kalifikazioa: Osatu gabea.)
[Ikaslearen erantzuna 3:] A terminala polo negatiboarekin lotu ez delako. (kalifikazioa: Zuzena.)
[Ikaslearen erantzuna 4:] A terminala polo negatiboarekin lotu delako (kalifikazioa: Kontraesana.)
```

Beste ikuspuntu batetik, esaldi pare baten antzekotasun maila identifikatzeko gai diren sistemak garatzeko asmoz Agirre *et al.* autoreek *STS* ataza plazaratu zuten (ingelesez *Semantic Textual Similarity*). Ataza hau aktibo egon da 2012. urtetik gaur egunera arte, eta komunitate zientifikoan interes handia piztu du. Ataza hau *hezkuntzarekin lotutako HP*arekin ere zeharo lotuta dago, gainera, autoreek ataza horren aldaera bat sortu zuten *iSTS* izenekoa (ingelesez *Interpretable Semantic Textual Similarity*). Azken ataza honetan, antzekotasun balioa ez ezik, esaldi parearen erlazioak ere identifikatu behar dira (Agirre *et al.*, 2016). 4 irudian *iSTS* atazaren egin beharrekoak eskematikoki azaltzen dira. Esaldi parearen erlazioak identifikatu eta sailkatu ondoren, sistemek posible dute antzekotasun eta desberdintasun hauekin osatutako berrelikadura lagungarria bueltatzea ikasleari. (Lopez-Gazpio *et al.*, 2016) lanean autoreak hasi dira *iSTS*ren eragina hezkuntzan aztertzen.

4 Irudia: *iSTS* atazan antzekotasun balioa ez ezik, esaldi parearen arteko loturak ere identifikatu eta sailkatu behar dituzte sistemek.



Gaur egun, pil-pilean dagoen beste datu-base batek *SNLI* du izena (ingelesez *Stanford Natural Language Inference*). Stanford unibertsitateak plazaratutako datu-base erraldoi bat da, esaldi pareak eta dagozkien loturen anotazioak biltzen dituena. Hain zuzen ere, artikulu honetan *SNLI* datu-basea erabiltzen duen neurona-sare konplexu bat aztertuko dugu (ikusi 3.2 atala). Honetaz gain, berau sortzeko motibazioa azaldu, inplementazioan sakondu eta *EB* desberdinak erabiliz lortzen ditugun emaitzak eztabaidatuko ditugu, hauek artearen egoeran lortu diren emaitzekin alderatuz. 5 irudian *SNLI* datu-basearen adibide bat ikus daiteke.

5 Irudia: *SNLI* datu-basearen itxurako adibide bat. *SNLI* datu-baseak esaldi pareak lotura mota batekin erlazionatzen ditu.

```
[Erreferentzia 1:] A terminala polo positiboarekin lotu da.
[Erreferentzia 2:] A terminala polo negatiboarekin lotu da.
[Lotura motak aukeran]: Kontraesana, baliokidetasuna edo ez bata eta ez bestea.
```

3 Ikerketaren muina

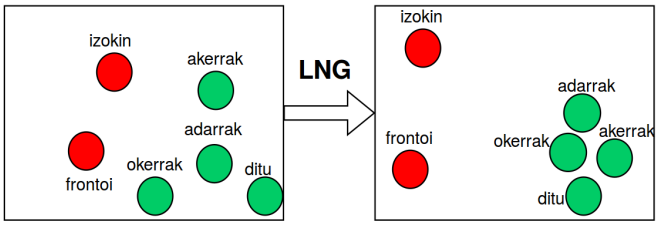
3.1 Bektore-espazio konbinaketak eta antzekotasun ataza

Aipatutako neurona-sareek hitz bakoitzari bi bektore esleituko dizkio; hitz moduan dituen ezaugarri semantikoak gordetzen dituena (W espazioan), eta testuinguruko hitz moduan dituen ezaugarriak dituena (C espazioan). 1 ataleko nomenklaturari jarraiki, bi espazioak $V \times d$ tamainakoak dira. Gauzak horrela,

*word2vec*ek corpus bat jasotzen du, eta bere xedea w hitz *EB*en eta c azken horien testuinguru *EB*ak hurbiltzea da.

*Word2vec*ek esaldiak irakurri ahala eguneratuko ditu hitzen *EB*ak; hots, hitz baten esanahia bere momentuko testuinguruaren arabera eguneratzen joango da. Demagun *word2vec* corpusa prozesatzen dabilela *"akerrak* ***adarrak*** *okerrak ditu"* esaldia jasotzen duela, eta momentu horretan *adarrak* hitzaren esanahia kalkulatu behar duela. Bada, alde batetik, *adarrak* hitzak Wn daukan *EB*a eta bere testuinguruko hitzek (*akerrak*, *okerrak* eta *ditu*) Cn duten *EB*ak hurbildu (berdez 6 irudian); bestetik, testuinguru horretan agertzen ez diren ausazko k lagin hartuko ditu W espaziotik (*izokin* eta *frontoi*, esaterako) eta urrundu (gorriz 6 irudian) egingo ditu. Azken teknika horri laginketa negatiboa (LNG) deritzo, eta optimizazioa azkartzeko modu oso efektiboa da. Hala, *akerrak* hitza corpusean berriro agertzerakoan, momentuko testuinguruko hitzekin eta lagin negatiboekin eguneratuko du. W eta C bereizita daude, eta M matrizea modu inplizituan erabiltzen da.

6 Irudia: Hitzen eta testuinguruen esanahien eguneraketa *word2vec*-en



Bada, *word2vec*en jatorrizko bertsioak testu corpusak soilik prozesatzen dituz, eta, ondorioz, erauzitako *EB*en esanahia testuko informazio semantikoaren izaerak determinatzen du. Hala ere, ezagutza-baseetako eta testu corpusetako informazio semantikoa osagarria dela jakinik, bi iturriak konbinatu daitezke. Hala, hainbat ikerlarik testu corpuseko *EB*ak ezagutza-baseetako informazioarekin aberastu dituzte ((Bollegala *et al.*, 2015), (Faruqui *et al.*, 2015)), eta antzekotasun-atazetan testu hutsarekin baino emaitza nahiko hobeak lortu. Teknika horiek aberaste-metodo legez izendatuko ditugu.

Aberaste-metodoekin alderatuta, testu eta ezagutza-base espazio bereizien konbinaketekin ((Goikoetxea *et al.*, 2016)) emaitzak ikaragarri hobetu dira antzekotasun atazan. Bada, Goikoetxea *et al.* autoreek hainbat bektore-espazio konbinaketa sinple egin, eta ingelesezko zazpi antzekotasun urre-patroitan ebaluatu zituzten: RG Rubenstein eta Goodenough (1965)), SimLex999 (SL) Hill *et al.* (2014), MTURK287 (MTU) Radinsky *et al.* (2011), MEN Bruni *et al.* (2014) eta WordSim353 (WS) Gabrilovich eta Markovitch (2007). Bada, RTE atazarako Goikoetxea *et al.* autoreen espazioen konbinaketetatik bi soilik erabiliko ditugu: testu corpusen eta ezagutza-baseetako pseudo-corpusen (Goikoetxea *et al.*, 2015) konbinaketa (COR), eta testu corpusen eta WordNet *EB*en kateaketa (KAT).

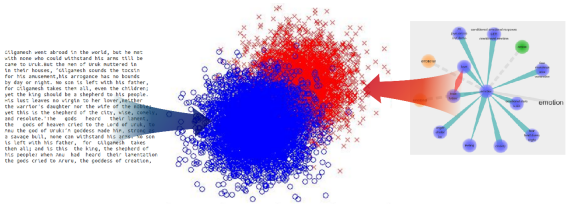
WordNet pseudo-corpus deritzoguna, WordNet (Miller, 1995) ezagutza-basean ausazko ibilbideak aplikatuta sortutako corpusa da, eta azken horren *EB*ek WordNeten erlazioak kodetzen dituzte. Hori horrela, COR konbinaketak testu hutsa eta WordNet pseudo-corpusak nahasten ditu, eta bere *EB*ak *word2vec* bidez erauzten. KATen, ordean, testu corpus eta WordNet pseudo-corpus bereizietatik erauzitako *EB*ak kateatzen dira. 1 taulan COR eta KAT Faruqui *et al.* autoreen metodoarekin (FAR) alderatzen dira.

1 Taula: FAR metodoaren, eta COR eta CAT konbinaketen Spearman korrelazioak, zazpi antzakotasun urre-patroitan

	RG	SL	MTU	MEN	WS
FAR	79.8	51.1	69.3	65.4	63.7
COR	83.0	47.9	67.9	**80.5**	**75.3**
KAT	**84.2**	**52.5**	**71.7**	78.7	71.4

1 taulako emaitzei so, KAT eta COR konbinaketak FAR metodoari alde handiagatik gailentzen zaiela ikusten da. Gauzak horrela, espazio bereizien konbinaketek aberasketa-metodoak baino *EB* osoagoak lortzen dituztela frogatzen da, eta, horregatik, hain zuzen, artikulu honetako RTE atazan KAT eta COR *EB*ak erabiliko ditugu sistemako sarrera legez.

7 Irudia: Testu eta WordNet bektore-espazio bereizien kobinaketak

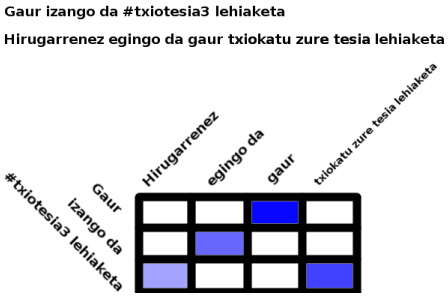


3.2 Testu-loturen azterketa *EB*ak oinarri hartuz

Hezkuntzarekin lotutako *HP*ri heltzeko, Parikh *et al.* autoreek garatutako sistema batean oinarrituko gara. Sistema honek 2.2 atalean deskribaturiko *SNLI* datu-basearekin lan egiten du (ikusi 5. irudia). Sistemaren motibazioa oso sinplea, baina aldi berean sendoa da: esaldi luze eta konplexuen arteko lotura logikoak identifikatzeko esaldi pareko segmentuen loturei erreparatzen baitie lehenbizi; eta, ondoren, lotura partzial hauen araberako kalifikazioa itzultzen du. Nolabait, 2.2 atalean deskribaturiko *iSTS* ataza inplizituki egikaritzen du (ikusi 4. irudia).

Sistema hiru azpiataletan antolatuta dagoen neurona-sare konplexuen konbinazioan oinarritzen da, jarraian deskribatzen ditugu atal hauek guztiak: lehen azpiatalean, *atentzioa* deritzona, sarrerako esaldi parea irakurtzen da eta hitz bakoitza dagokion *EB*arekin lotzen da. Ondoren, hitzen *EB*ak konbinatu egiten dira *neural attention* delako teknika baten bitartez (eragiketa aljebraikoetan oinarrituta). Honen bitartez, esaldi parean dauden hitz guztiak parekatzen dira antzekotasunaren arabera 8. irudian ikus daitekeen itxurako matrize bat eratuz.

8 Irudia: Operazio aljebraikoak erabiliz hitzen *EB*ak konbinatu egiten dira esaldi parearen antzekotasunak esplizitu eginez.



Behin antzekotasun matrize hau daukagula esaldi batek beste esaldiko segmentuengan duen eragina aztertu daiteke. Bertan, bigarren esaldiak (lerroen baturak) lehen esaldiko hitzekiko (zutabeetan dauden hitzekiko) lotura aztertzen da, eta, baita, lehen esaldiak bigarren esaldiaren hitzekiko lotura ere. Lan honi esaldi batek bestearekiko duen *projekzioa* deritzo, eta erauzitako informazioa sistemaren bigarren atalak erabiliko du: *Konparatu* azpiatalak.

Bigarren azpiatalak, alde batetik, erauzitako proiekzio guztiak, eta, bestetik, jatorrizko hitzen *EB*ak konparatu egiten ditu neurona-sare bat erabiliz. Era honetan, bi esaldien artean dauden loturak kuantifikatu egiten ditu, loturaren indarra kontuan hartuz: lotura garrantzitsuei indar gehiago ematen ikasten du, eta, aldiz, lotura ahulei indar txikiagoa.

Sistemaren azken atalean, *batu* deritzona, esaldi parearen lotura guztiak bildu egiten dira, eta bukaerako emaitza itzultzen da. Bukaerako emaitza hau itzultzeko ere beste neurona-sare bat erabiltzen da. Sare honek adieraziko du esaldi parearen artean lotura, aurkakotasuna edo ez bata eta ez bestea antzeman duen. Sistema osoak *SNLI* datu-basean entrenatzeko egunak har ditzake, milaka eta milaka esaldi pare baitaude bertan. 2. taulan deskribatutako sistema erabilita lorturiko emaitzak plazaratzen ditugu. 3. taulan ikus daitekeen modura zehaztasun maila altua lortzen dugu sistema honen eta 3.1 atalean deskribatutako *EB*en bitartez. Hala eta guztiz ere, badira beste autore batzuk emaitza hobeak lortu dituztenak arkitektura konplexuagoak erabilita. Kasu, Wang *et al.* autoreek lan honi esaldien egitura

2 Taula: Hitz *EB* desberdinak erabilita *SNLI* datu-basean lorturiko emaitzak. Emaitzek datu-basearen gaineko zehaztasuna islatzen dute ehunekotan, honela: kalifikazio zuzena esleitu zaien pare kopurua pare kopuru osoarekin zatituta.

	Ikasketa-multzoa	Ebaluazio-multzoa
3.1 ataleko *COR EB*ekin	%89.5	%84.4
3.1 ataleko *KAT EB*ekin	%85.2	%83.4

sintaktikoa ere kontuan hartzen duen neurona-sare gehigarri bat txertatu zioten %93 eta %89 ko zehaztasuna lortuz ikasketa-multzo eta ebaluazio-multzoan hurrenez hurren[1].

4 Ondorioak

Artikulu honetan bi ikerketa-lerro jorratu ditugu: alde batetik, semantika distribuzionala eta bektore espazioen konbinaketak, eta, bestetik, testu-loturak eta hauen aplikazioa hezkuntzan. Hasieran, hitzak *EB*tan kodetzeko prozedurak aztertu ditugu eta espazio desberdinen konbinaketak emaitzak hobetzera eramaten gaituela ikusi dugu. Gainera, hitzak *EB*tan kodetuta izateak ematen dizkigun abantailak aipatu ditugu. Izan ere, *EB*ekin marko matematiko batean lan egin dezakegu, eta, era honetan, adimen artifizialeko teknikak erabili hitzak konbinatzeko eta esaldi mailara salto egiteko. Esaldi mailara salto egitean, beste mota bateko atazak ere burutu ditzakegula ikusi dugu, hala nola: esaldien antzekotasuna edota testu-loturen analisia. Azken hau egikaritzen duen sistema bat ere deskribatu dugu eta bere emaitzak plazaratu ditugu. Oro har, semantika distribuzionala eta hezkuntzari lotutako *HP* pil-pilean dauden atazak dira egun, eta, artearen egoera oso azkar aldatzen ari da uneoro.

5 Etorkizunerako planteatzen den norabidea

Ezagutza-baseen eta testu corpusen informazio semantikoaren osagarritasuna hainbat modutan konbinatu baditugu ere, artikulu honetako esperimentu oro ingelesez soilik egin dira. Hala, aipatutako bektore-espazioen konbinaketez gain, ezagutza-baseen hizkuntzarteko informazioa ustiatzeak potentzial handia dauka. Izan ere, ezagutza-base batzuetan (*WordNet* eta *Wikipedia*, esaterako) kontzeptuak hizkuntzen independenteak dira, eta kontzeptu horiek hizkuntzen arteko zubi legez erabili daitezke (*txakur* eta *dog* hitzek, esaterako, kontzeptu bera partekatzen dute), eta hizkuntzarteko lotura horiek oso baliagarriak izan daitezke, *SNLI*n zein beste hainbat atazatan (itzulpen automatikoan, sentimenduen analisian, antzekotasunean...).

*HP*an azpimarratu daitekeen beste ondorioetako bat ondorengoa da: zenbat eta teknika gehiago modu eraginkorrean konbinatu, are eta emaitza hobeak lortu ohi direla. Hori dela eta, argi dago 3.1 eta 3.2 ataletan deskribatutako sistemei teknika gehigarriak integratzeko estrategia berrien bila jarraitu behar dugula lanean.

Erreferentziak

Agirre, Eneko, Daniel Cer, Mona Diab, eta Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 385–393, Montréal, Canada. Association for Computational Linguistics.

——, Aitor Gonzalez-Agirre, Inigo Lopez-Gazpio, Montse Maritxalar, German Rigau, eta Larraitz Uria. 2016. Semeval-2016 task 2: Interpretable semantic textual similarity. *Proceedings of SemEval* 512–524.

Bollegala, Danushka, Alsuhaibani Mohammed, Takanori Maehara, eta Ken-ichi Kawarabayashi. 2015. Joint word representation learning using a corpus and a semantic lexicon. *arXiv preprint arXiv:1511.06438* .

Bruni, Elia, Nam-Khanh Tran, eta Marco Baroni. 2014. Multimodal distributional semantics. *JAIR* 49.1–47.

[1] `http://nlp.stanford.edu/projects/snli/` webgunean *SNLI*n ebaluatu diren sistema guztiak atzitu daitezke.n ebaluatu diren sistema guztiak atzitu daitezke.

Dzikovska, Myroslava, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, eta Hoa Trang Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, eta Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*, 1606–1615.

Gabrilovich, Evgeniy, eta Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *IJCAI*, volume 7, 1606–1611.

Goikoetxea, Josu, Eneko Agirre, eta Aitor Soroa. 2016. Single or multiple? combining word representations independently learned from text and wordnet. In *AAAI*, 2608–2614.

——, Aitor Soroa, eta Eneko Agirre. 2015. Random Walks and Neural Network Language Models on Knowledge Bases. In *Proceedings of NAACL-HLT*, 1434–1439.

Harris, Zellig S. 1954. Distributional structure. *Word* .

Hill, Felix, Roi Reichart, eta Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456* .

Leacock, Claudia, eta Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities* 37.389–405.

Lopez-Gazpio, I, M Maritxalar, A Gonzalez-Agirre, G Rigau, L Uria, eta E Agirre. 2016. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems* .

McNamara, Danielle S, Scott A Crossley, Rod D Roscoe, Laura K Allen, eta Jianmin Dai. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing* 23.35–59.

Mikolov, Tomas, Kai Chen, Greg Corrado, eta Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Miller, George A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38.39–41.

Nielsen, Rodney D, Wayne H Ward, eta James H Martin. 2008. Learning to assess low-level conceptual understanding. In *Flairs conference*, 427–432.

Parikh, Ankur, Oscar Täckström, Dipanjan Das, eta Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2249–2255, Austin, Texas. Association for Computational Linguistics.

Radinsky, Kira, Eugene Agichtein, Evgeniy Gabrilovich, eta Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of WWW*, 337–346.

Rubenstein, Herbert, eta John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8.627–633.

Wang, Zhiguo, Wael Hamza, eta Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814* .

Zechner, Klaus, Keelan Evanini, Lei Chen, Shasha Xie, Wenting Xiong, Fei Huang, Jana Sukkarieh, eta Miao Chen, 2015. Computer-implemented systems and methods for content scoring of spoken responses. US Patent 9,218,339.

6 Eskerrak eta oharrak

Josu Goikoetxearen eta Iñigo Lopez-Gazpioren lanak garatzen ari diren doktoretza-tesietan oinarrituta daude. Aipatutako tesiak burutzeko Josuk UPV/EHU Euskara Errektoreordetzako beka dauka, eta, Iñigok ministerioak emandako FPU beka. Azkenik, IXA taldea eskertu nahiko genuke artikulu hau aurrera eramateko baliabideengatik.