EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

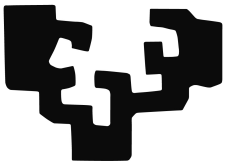
Doktorego-tesia

Hitzen arteko antzekotasuna: ezagutza-baseetan oinarritutako tekniken ekarpenak

Josu Goikoetxea Salutregi

2018

eman ta zabal zazu



EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

Hitzen arteko antzekotasuna: ezagutza-baseetan oinarritutako tekniken ekarpenak

Josu Goikoetxea Salutregik Eneko Agirre Bengoaren eta Aitor Soroa Etxaberen zuzendaritzapean egindako tesi-txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua.

Donostian, 2018ko ekainean.

Orduan, zertarako balio dute hitzek eta eleek? (...)
Hitza, bilatzera bultzatzen zaituelako da baliagarri (...).
Zentzuaren bila bultzatzen zaitu,
baina ez da bera zuk ikusi gura duzuna.

Rumi.

Eskerrak

Lehenik eta behin, nire familiari eskerrak eman nahi dizkiot. Aurreko ibilbide profesionala eta pertsonala alde batera utzi, eta abentura berri honetan sartzeko erabakiaren alde egin zenutelako, eta hasieratik nigan sinistu duzuelako. Eskerrik asko, Valentina, Iban eta Eider! Eider, bai, zu ere talde horretan sartzen zaitut!

Hurrengo eskerrak nire bi zuzendarientzako dira, Aitor eta Eneko! Nire "neurak" jasatearren, 300 dimentsiotako espazioetan nire iparrorratzak izatearren, eta beti hor egotearren. Tesi-lan honetan ikasitako guztia zuiei zor dizuet! Esan gabe doa, IXA Taldeko kide guztiei ere sakonki eskerturik nagoela. Leku askotan ibili naiz lanean azken 15 urteotan, baina, inon ez dut halako talderik aurkitu... Eskerrik asko pintxopote, sagardotegi, menditxango, "kepasada" eta beste plan guztiegatik, baina, batez ere, ingenieritza gordinenetik etorritako ardi galdu honi lan egiteko beste modu bat erakustearren. Itzelak zarete!

Azken bost urteotan hainbat lagunek eman didazue sostengua kontu honekin, Gernikan, Innsbrücken, Donostian, Bilbon, Iruñean eta Navaridasen. Ezin denen izenak zerrendatu, luzerako emango bailuke... Eskerrik asko zeuei ere, badakizue nortzuk zareten! Eskerrak baita ere, azken urteotan nire bizitzan gurutzatu zareten guztioi, zuzenean ez bada ere, denok jarri baituzue zuen aletxoa tesi-lan honetan.

Umea nintzenean, nire aitona zenaren ahotik zera entzuten nuen: "euskeriek ez deu ezetako balixo!". Nire amona zenak, ordea, zera zioen: "erderiek ez deu ezetako balixo!". Aitonaren berben azpian tristura sumatzen nuen

beti, eta, amonarenen azpian, haserrea... Tesi-lan hau, besteak beste, umetatik barneratutako ikuspegi eta sentipen kontraesankor horiek bateratzeko tresna izan da niretzat. Hein batean, beraz, nire aitonari eta amonari ere oso eskerturik nago. Aurrean izango banitu, zera esango nieke: "euskerie ta erderak, danak diez baliotsuek!".

Finalmente, querría agradecer también a mi "otra familia"; MariBel y Jose Luis, Maria *metalhead* Izquierdo, Igor, Amaia, Itzi, Marga, MariBel, Maria, Itxaso, Arantza y a los demás (sois demasiados para nombraros a todos). Vuestro apoyo incondicional ha sido y sigue siendo esencial en este viaje. Gracias por hacer de faro en las tormentas, por vuestra comprensión y sinceridad, por creer en mí a pesar de mis dudas y exigencias, y, sobre todo, por vuestra humanidad. Eskerrik asko!

Tristuran eta pozean, zalantzan eta ziurtasunean, beldurrean eta ausardian, nire deabrutxo pertsonalen eta aingeruen lekuko eta ispilu izan zarete, nire bidelagun!

Eskerrak denoi, bihotz-bihotzez!

Esker instituzionalak

Eskerrak Euskal Herriko Unibertsitateko Euskara Errektoreordetzari, tesi-lan hau euskaraz egiteko emandako bekarengatik.

Laburpena

Eredu konputazionalekin sortutako hitzen errepresentazio semantikoak gakoa dira hizkuntzaren prozesamenduko hainbat atazatan, eta errepresentazio horien kalitatea ebaluatzeko hitzen arteko antzekotasuna erabiltzen da. Antzekotasun-ataza hizkuntzaren prozesamenduaren alorrean kokatzen da, lexiko-semantikan, eta, hurrengo urratsak ditu: lehenik, hitzen arteko antzekotasuna hitzen errepresentazioen bidez kalkulatzen da; ondoren, antzekotasun hori gizakien antzekotasun-irizpideekin konparatzen da. Eredu konputazionalaren emaitzak zenbat eta gizakion irizpideetatik hurbilago egon, orduan eta kalitate hobea izango dute hitzen errepresentazioek. Lan honetan antzekotasunaren kasu orokorragoarekin ere lan egin dugu, ahaidetasunarekin.

Hitzen errepresentazioan testu-corpusetan oinarritutako metodoak eta ezagutza-baseetan oinarritutakoak daude. Aurreneko familian hainbat eredu daude, baina, lan honetan neurona-sareetan oinarritutakoak erabili ditugu. Metodo horiek hitzen esanahiak testuetako hitz-testuinguru agerkidetzen bidez inferitzen dituzte eta bektore-espazio trinko batean kodetzen. Bigarren familiakoen artean, ezagutza-baseak grafoak balira bezala tratatzen dituztenez baliatu gara, azken horien informazio estrukturala bere osotasuenan ustiatuz.

Tesi-lan honen xedea antzekotasun-atazako emaitzak hobetzea da, eta, azken hori hitzen errepresentazio semantiko hobeak erdiesteko teknikez burutuko dugu. Gure hipotesi nagusia testu-corpusetako eta ezagutza-baseetako informazioa desberdina eta osagarria dela da. Gure aburuz, bi iturri horiek konbinatuz gero hitzen errepresentazioen arteko antzekotasun-emaitzak ho-

betuko dira, eta, ondorioz, errepresentazio hobeak izango ditugu. Hipotesi hori, gainera, elearteko erlazioetara hedatu dugu.

Tesi-lan honen bitartez aurreko paragrafoko hipotesiak frogatu ditugu, eta egindako ekarpenak hurrengo hirurak dira: (1) ausazko ibilbideen metodo batekin ezagutza-baseetako informazio estrukturala corpus batean kodetzea, eta azken horren hitzen errepresentazio semantikoak kalkulatzea; (2) testuko eta ezagutza-baseetako informazio semantikoa konbinatzeko hainbat metodo eta errepresentazio hibrido proposatzea; (3) aurretik proposatutako guztiak elearteko erlazioetan aplikatzea.

Aipatuako metodo eta konbinaketa oro antzekotasun-atazan ebaluatu ditugu, beren emaitzak artearen egoerako metodo baliokideekin konparatuz. Gure proposamenek antzekotasun-atazako artearen egoera berdindu edo gainditu dute, eta gure hipotesiak betetzen direla ondorioztatu dugu.

Gaien aurkibidea

1 Sarrera

Hitzen arteko antzekotasun semantikoa hizkuntzaren prozesamenduko hainbat atazaren muinean dago, hala nola, informazio erauzketan, hitzen adiera desanbiguazioan, dokumentuen gai-detekzioan, itzulpen automatikoan edota testu-loturan. Hala, antzekotasuna erreproduzitzeko gaitasunak eredu konputazional batek kalkulatutako hitzen errepresentazio semantikoen kalitatea ebaluatzeko balio du. Zehazki, hitzen errepresentazioen antzekotasunak zenbat eta giza irizpideetatik hurbilago egon, orduan eta hobeto modelatuak egongo dira hitzen esanahiak.

Tesi-lan honen xede nagusia hitzen arteko antzekotasun-atazako emaitzak hobetzea da, eta, horretarako, hitzen errepresentazio semantiko hobeak erdiesteko teknikak proposatu ditugu. Horiek guztiak lortze aldera, lexiko-semantikan dauden bi alor nagusietako teknikez baliatu gara: testu-corpusetan oinarritutakoak eta ezagutza-baseetan oinarritutakoak. Izan ere, alor horietako teknikek informazio semantiko desberdinetako baliabideak ustiatzen dituzte, hitzen arteko antzekotasun-erlazioen ñabardura desberdinak jasoz.

Lan honetan testu-corpusetako eta ezagutza-baseetako informazio semantikoa konbinatzeko hainbat metodo proposatu ditugu, baita azken hori dagozkien hitzen errepresentazio hibridoak ere. Aipatutako konbinaketa horiek elearteko espazioetara ere hedatu ditugu, elearteko antzekotasunaren nondik norakoak esploratuz.

1.1 Antzekotasuna

Antzekotasuna hainbat hausnarketa filosofikoren erdian egon da antzinatetik, gizakion kontzientzia eta gogamena ulertzeko giltzarri moduan ulertu izan baita beti. Bada, antzekotasunari ikuspegi psikologikotik heldu zion aurrenekoa Aristotelesena izan zen, antzinako Grezian. Platonen inspiratuta, Aristotelesek prozesu psikologikoak aztertu zituen, K.a. 300 inguruan hurrengo lau legeak proposatuz (Boeree, 2000):

- Jarraitasuna: hurbiltasun espaziala edo tenporala duten gertaerak gogamenean asoziatuta egoteko joera daukate.
- Maiztasuna: gertaera biren jazoera kopurua gertaera horien asoziazioaren proportzionala da.
- Antzekotasuna: gertakari baten pentsamenduek antzeko pentsamendua sorrarazteko joera daukate.
- Kontrastea: gertakari baten pentsamenduek aurkako pentsamendua sorrarazteko joera daukate.

Aristotelesek lege horiek gizakion sen onaren edo zentzuaren (*common sense*) oinarri legez deskribatu zituen[1] eta Asozianismoaren (Burnham, 1888) abiapuntua izan ziren. Azken paradigma hori hurrengoan oinarritzen da: gogamena elementu kontzeptualen multzoa da, eta azken horiek beren arteko asoziazioen bitartez antolatzen dira. Bada, gure ikerketa-ildoan filosofo greziarrak proposatutako legeetako baten gainean soilik jarriko dugu arreta, antzekotasunean.

Hizkuntzaren prozesamenduan antzekotasuna ulertzeko psikologia kognitiboan oinarritutako ikuspegia jarraitzen dela azpimarratzea garrantzitsua da, eta ez Asozianismoan (edo azken horretan oinarritutako teoriaren batean). Izan ere, XX. mendearen erdialdean iraultza kognitiboa deiturikoa mugimendu intelektualak (Miller, 1956; Skinner, 1957/2014; Neisser, 1967/2014) kognizioa ezarri zuen prozesu psikologikoak ulertzeko paradigma moduan. Iraultza kognitiboaren ondoren, Asozianismoaren eragina gutxitu egin da

[1] Esplizituki aipatuko ez baditugu ere, lege horiek hainbat bider gogoratuko ditugu tesi-lan honetan zehar. Izan ere, egungo ikasketa automatikoko eta ikasketa sakoneko oinarrizko asumitzeak dira (Wang *et al.*, 2017), hots, tesi-lan honetan erabilitako metodoen asumitzeak (inplizituak).

psikologian eta psikolinguistikan. Asozianismoa paradigma nagusia ez bada ere, oraindik bizirik dirau maila teorikoan (Shanks, 2010).

Psikologia kognitiboaren ikuspegitik, antzekotasuna gizakion berezko gaitasuna da, modurik naturalenean darabilgu egunero, inongo zailtasun barik. Objektuak behin eta berriro konparatzen ditugu. Hainbat antzekotasun irizpide erabiltzen ditugu eguneroko gertakariak eta entitateak sailkatzeko, errealitatea deritzogun hau egituratzeko. Irizpide horien aniztasuna kontuan izanik, antzekotasuna dirudiena baino gaitasun konplexuagoa da, ez da gauza soila, elkarrekin erlazionatutako prozesu anitzez osatua baizik (Vosniadou and Ortony, 1989). Bada, antzekotasun irizpideen sistema horiek garatzea haurtzaroaren hasierako lorpen kognitibo nagusietakoa da (Vosniadou and Ortony, 1989). Willard Van Orman Quine filosofoak gizakion ikasketa-prozesuaren inguruan hainbat ekarpen egin zituen, eta prozesu hori guztia "antzekotasun sen primitiboan" datzala zioen. Quinek *original sim* hitz-jokoa ere erabiltzen zuen; hau da, ingelesez esplizituki *antzekotasun originala* esan gura badu ere, inplizituki bekatu originalari egiten dio erreferentzia. Hala, filosofoaren hitzetan:

> *Antzekotasuna oinarrizko kontua da ikasketan, jakintzan eta pentsamenduan. Izan ere, gure antzekotasunerako senak errealitateko gauzak mota desberdinetan ordenatzen ahalbidetzen digu (...). Zentzuzkotzat duguna hurrengoen mende dago; zirkunstantzien antzekotasuna, eta, antzeko kausek antzeko efektuak izango dituztela itxaroteko joera. (Quine, 1969)*

Amos Tversky psikologo kognitiboak ere antzekotasunaren izaera sakon aztertu zuen. Aurreko paragrafoetan esandakoaren ildo berean, zera zioen:

> *Antzekotasunak (...) antolakuntza-printzipio legez balio du, gizabanakoek objektuak klasifikatzeko, kontzeptuak osatzeko, eta orokortzeak burutzeko. (Tversky, 1977)*

1.1. irudiak haur bat irudikatzen du, jolasean murgilduta. Jolasak, baina, ludikotasunetik askoz haratago doaz, munduari zentzua ematen ikasten dute beren jolasen bitartez. Errealitatearen ikasketa-prozesu horretan antzekotasuna gaitasun kognitibo garrantzitsua da. Izan ere, antzekotasuna gizakion sortzetiko gaitasun kognitiboa da, ikasketa-prozesuetan eragile zinez aktiboa haurtzaro goiztiarrenetik. Bada, antzekotasunak izaera egituratzailea dauka,

1.1 irudia – Haurrek munduari zentzua ematen ikasten dute beren jolasen bitartez. Errealitatearen ikasketa-prozesu horretan antzekotasuna gaitasun kognitibo garrantzitsua da. Iturria `https://goo.gl/Qx21oC`.

giza errealitatea hezur-mamitzen duen itsaski kognitibo ikusezina da; hots, *gure errealitatearen* funtsezko oinarrietako bat.

Gizakiontzat hain naturala dena, baina, makinetan erreproduzitzea ez da batere erraza. Gaitasun kognitibo horren ezaugarriak motibazio legez hartuta, makinek ere errealitatearen inguruko informazioa ordenatzea eta egituratzea nahi badugu, antzekotasun-irizpideak erakusteak berebiziko garrantzia du.

Hizkuntza eragile zinez aktiboa izanik haurren garapen kognitiboan (Vygotsky, 1980; Piaget, 1959/2005), antzekotasunaren ikuspegi kognitiboak hizkuntzalaritzan itzal handia du. Tesi-lan honi dagokionez, hizkuntzalaritzak hitzen arteko antzekotasunari buruz duen ikuspegia interesatzen zaigu, eta, azken horren aburuz, antzekotasuna eta hitzen errepresentazio semantikoak eskutik doaz, ezin bata bestea barik ulertu.

Hitzak errepresentatzeko eredu konputazionalak deskribatu aurretik, baina, hurrengo atalean antzekotasunaren inguruan ñabardura bat egingo dugu.

1.2 irudia – Otsoak txakurrarekin eta ilargiarekin antzekotasun eta ahaidetasun erlazioak ditu, hurrenez hurren.

Antzekotasuna eta ahaidetasuna bereizten

Orain arte antzekotasuna denominazioa modu orokorrean erabili badugu ere, zientzia kognitiboetan zinezko antzekotasunaren eta asoziaziozko antzekotasunaren arteko desberdintasuna ezaguna da aspaldidanik (Tversky, 1977). Lan honetan zinezko antzekotasunari antzekotasun deituko diogu eta asoziaziozkoari ahaidetasun.

1.2. irudiko adibideak antzekotasun eta ahaidetasun erlazioak deskribatzen ditu. Alde batetik, otsoak eta txakurrak antzekotasun handia erakusten dute, biak kanidoen familiako animaliak dira, eta familia horren ezaugarriak guztiak partekatzen dituzte (azeriarekin batera). Otsoaren eta ilargiaren arteko erlazioa, ordea, ez da antzekotasunezkoa inondik inora ere, ez baitute inongo ezaugarririk partekatzen. Hala ere, mendebaldeko kulturan bederen, otsoa-ilargiari-uluka-egiten irudia denoi egiten zaigu ezagun. Azken hori, hain zuzen, asoziaziozko erlazioa da, ahaidetasuna.

Gauzak horrela, tesi honetan Budanitsky and Hirst-en (2006) interpretazioa erabiliko dugu antzekotasuna eta ahaidetasuna bereizteko:

- Antzekotasuna: hitzen arteko sinonimia (*okela/haragi*), hiponimia/hiperonimia (*kolore/urdin*).

- Ahaidetasuna: aurreko erlazio semantikoez gain, meronimia (*oin/behatz*), antonimia (*argi/ilun*), asoziazio funtzionalak (*arropa/armairu*) eta bestelako ezohiko erlazioak.

Ahaidetasuna, beraz, oso orokorra da, eta antzekotasuna ahaidetasun kasu berezia. Hizkuntzaren prozesamenduan bereizketa hori egitea garrantzitsua da, aplikazioaren arabera bata ala bestea izango baizaigu baliagarria. Esaterako, dokumentuen gai-detekzioan asoziatutako hitzak antzekoak baino garrantzitsuagoak dira: *begi* eta *betaurreko* erlazionatuta daudela jakitea informatiboagoa da gai-detekzioan *hitz* eta *berba* sinonimoak direla jakitea baino. Alderantziz, gure errepresentazioetan *hitz* eta *aho* ertsiki erlazionatuta badaude ere, itzulpen automatikoan *hitz* ez genuke ingeleseko *mouth* bezala itzuli beharko, *word* bezala baizik.

Esan bezala, antzekotasunaren edota ahaidetasunaren erreproduzitzeko, lehenik hitzen errepresentazio semantikoak behar ditugu. Hurrengo atalean tesi-lan honetan esanahia modelatzeko bi eredu konputazionalen sarrera egingo dugu, biak ere oso erabiliak hizkuntzaren prozesamenduan.

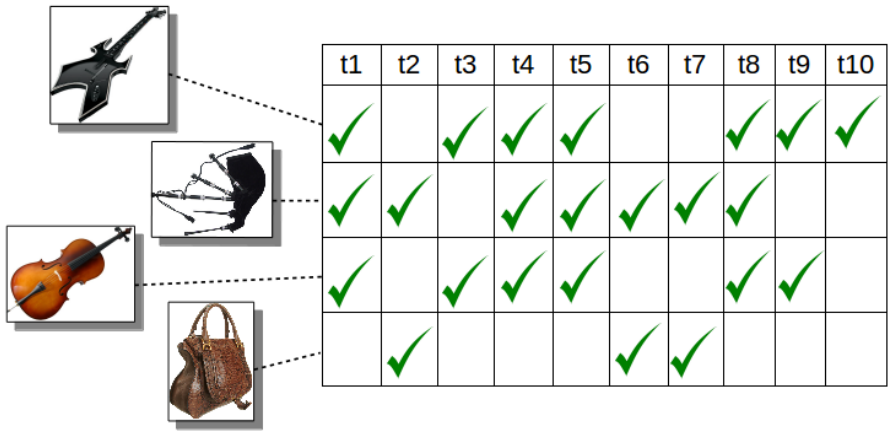
1.2 Kogniziotik konputaziora

Aurreko ataletan antzekotasun orokorra esanahiari ertsiki lotutako gaitasun kognitiboa dela aipatu dugu. Ildo horretan, Informazioaren Teoriaren alorrean Lin *et al.*-ek (1998) antzekotasun orokorraren definizio formala planteatzen du, hurrengo hiru giza intuizioetan oinarrituz:

- A eta B objektuen arteko antzekotasunak komunean daukatenarekin lotuta dago. Zenbat eta komunean gehiago eduki, orduan eta antzekoagoak izango dira.

- A eta B objektuen arteko antzekotasunak beraien arteko desberdintasunekin alderantzizko erlazioa dauka. Desberdintasuna areagotu ahala, antzekotasunak behera egingo du.

- A eta B berberak direnean, bien arteko antzekotasuna maximoa izango da



1.3 irudia – Lau terminoren errepresentazio semantikoak eredu distribuzional soil batean. Lerro bakoitzeko terminoak hamar tasunen bidez (t1-t10 tasunak) errepresentatzen dira.

Baina, nola gauzatu giza intuizio horiek eredu konputazional batean? Lehenik eta behin, hitzen errepresentazio semantiko bat osatu behar dugu. Azken hori burutzeko, hizkuntzaren prozesamenduan zein beste hainbat alorretan, hitzen ezaugarrietan oinarritzea estrategia oso eraginkorra eta zabaldua da; hots, termino baten tasun esanguratsuenak adierazpen batean laburbiltzea[2]. Gero, adierazpen horien tasunen arteko alderaketa kuantitatibo baten bitartez terminoen arteko antzekotasuna neur dezakegu. Gauzak horrela, aipatutako antzekotasunaren inguruko hiru giza intuizioak modu kuantitatiboan kalkulatzeko gai izango gara[3].

1.3. irudiko adibidean, lau terminoren errepresentazio distribuzionalak agertzen dira, gitarrarena, gaitarena, txeloarena eta poltsarena, hain zuzen. Zutabeek termino horien tasunak (*t1-t10* tasunak) adierazten dituzte[4], (terminoen esanahia definitzen dute) eta termino bakoitza hainbat tasunen bitar-

[2]Erabilitako metodoen arabera, esanahi baten tasunak latenteak zein konkretuak izan daitezke. Lan honetako errepresentazio guztien tasunak latenteak dira.

[3]Lin *et al.*-en (1998) hiru giza intuizioen bertsio kuantitatiboak hurrengoak dira: zenbat eta bat-etortze handiagoa tasunen artean, orduan eta antzekoagoak terminoak; zenbat eta bat-etortze txikiagoa, antzekotasun gutxiago; tasun berberak izanez gero, terminoak berdinak dira.

[4]Adibidean ez da beharrezkoa tasun zehatzak zeintzuk diren jakitea. Hala nola, tasunetako batzuk hurrengoak izan daitezke: material zurrunez egina, sokez egina, telaz egina, musika-tresna, e.a.

Hitz bikotea	Teilakatzea	Antzekotasuna
gitarra-gaita	4	0,4
gitarra-txeloa	6	0,6
gitarra-poltsa	0	0,0
gaita-txeloa	4	0,4
gaita-poltsa	3	0,3
txeloa-poltsa	0	0,0

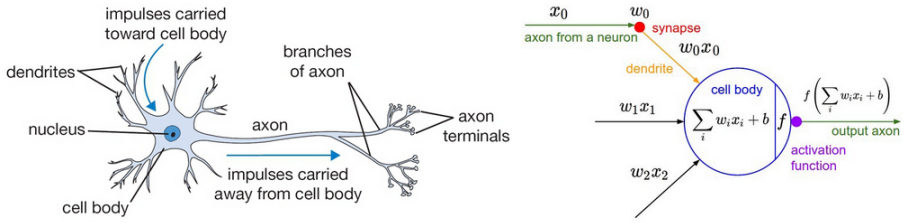
1.1 taula – 1.3. irudiko termino-bikote posible guztien teilakatzeen eta antzekotasunen taula. Antzekotasuna kalkulatzeko, 1.3. irudiko terminoen tasunen arteko teilakatzea zati terminoak definitzeko erabilitako tasun kopurua (hamar).

tez definitzen da. Demagun adibideko terminoen arteko antzekotasuna tasunen teilakatze kopuruarekin neurtuko dugula; zehazki, teilakatzea zati tasun kopuru maximoa (hamar). Beraz, antzekotasunaren balioa zerotik (inongo antzekotasunik gabeko terminoak) batera (antzekotasun maximoa, termino bera) joango da. Hala, 1.1. taulak tasunen teilakatzeak eta antzekotasun-balioak laburbiltzen ditu.

1.1. taulako balio kuantitatiboei begira, Lin *et al.*-en (1998) giza intuizioak modu errazean gauzatu ditugu. Bada, *gitarra-txeloa* parea da antzekotasun altuena duena, *gitarra-gaita* eta *gaita-txeloa* pareek jarraitzen diete, ondoren, *gaita-poltsa* dator, eta, azkenik, *gitarra-poltsa* eta *txeloa-poltsa* pareek ez dute inongo antzekotasunik.

Modu soilean bada ere, tasunen teilakatzean oinarritutako irizpide horrekin giza intuizioen hurbilketa gauzatu dugu, gizakion erantzunekin gutxi gorabehera bat datorrena. Esanahia eta antzekotasuna konputazionalki modelatzeko adibide honen ondoren, tesi-lan honetan erabilitako eredu konputazionalak aurkeztuko ditugu.

Lan honen abiapuntua hizkuntzaren prozesamenduko bi arlo nagusitako teknikak dira, hots, testu-corpusetan oinarritutako teknikak eta ezagutza-baseetan oinarritutakoak. Bi metodo horiek izaera semantiko desberdineko baliabide linguistikoekin lan egiten dutenez, antzekotasunean edota ahaidetasunean ekarpen desberdinak egiten dituzte. Hurrengo bi ataletan bi metodo horien nondik norakoen sarrera bat egingo dugu.



1.4 irudia – Pertzeptroia (Rosenblatt, 1958) neurona biologikoan inspiratutako eredu konputazionala da. Pertzeptroia egungo neurona-sare konplexuen oinarria da, eta, sarrerez (x_i), pisuez (w_i), aktibazio funtzio batez f eta irteeraz osatuta dago. Iturria `goo.gl/fhHECc`.

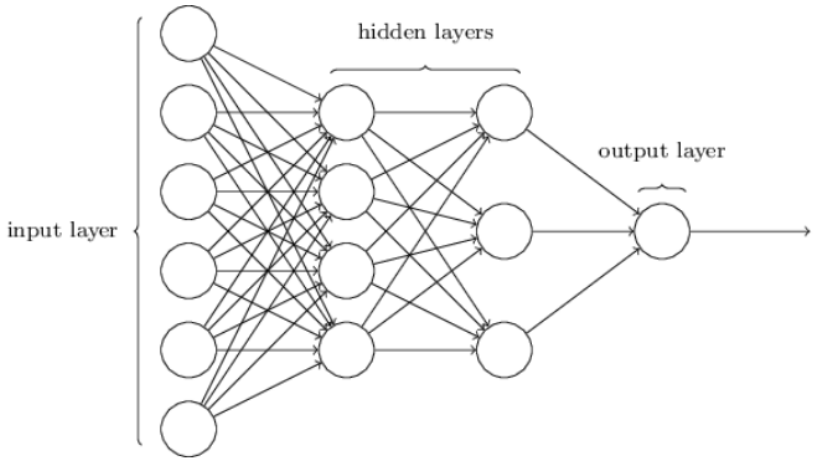
Corpusetan oinarritutako metodoak

Corpusetan oinarritutako metodoak Semantika Distribuzionalean oinarritzen dira, zehazki, hipotesi distribuzionalean (Harris, 1954): hots, hitzen esanahia beren testuinguruek definitzen dutela asumitzea (xehetasun gehiago 2.1. atalean). Azken hipotesi horri jarraiki, corpusetan oinarritutako metodoen artean bi familia nagusi agertu dira, kontaketetan oinarritutakoak eta iragarpenetan oinarritutakoak (ik. 2.1.2. atala). Tesi-lan honetan azkenengo familiako metodoak erabili ditugu, kontaketetan oinarritutakoak baino emaitza hobeak ematen baitituzte (Baroni *et al.*, 2014). Iragarpen metodoen artean, zehazki, hizkuntza-ereduetan oinarritutako neurona-sareetaz (Bengio *et al.*, 2003; Collobert and Weston, 2008; Mikolov *et al.*, 2013a) baliatu gara.

Aurreko paragrafoan esan bezala, tesi-lan honetan corpusetako hitzen errepresentazioak neurona-sareen bitartez erauzi ditugu. Azken horiek, XX. mendearen erdialdetik proposatu ziren (Hebb, 1949/2005; Rosenblatt, 1958), eta, jatorrian, neurona biologikoak informazioa prozesatzeko moduan inspiratu ziren. Aurreneko eredu konputazionala pertzeptroia (Rosenblatt, 1958) izan zen, neurona bakarraz osatutakoa. 1.4. irudian neurona biologikoaren eta pertzeptroiaren arteko baliokidetasunak erakusten ditu. Bada, neurona biologikoari jarraiki, pertzeptroiak, lehenik, sarrerak (x_i) jasotzen ditu, pisuak aplikatzen dizkie (w_i) eta (1.1) ekuazioko konbinaketa lineala aplikatzen die:

$$y = \sum_i w_i \cdot x_i + b_i \tag{1.1}$$

Ekuazio horretan y konbinaketa lineala da eta b_i bias terminoa. Ondoren,



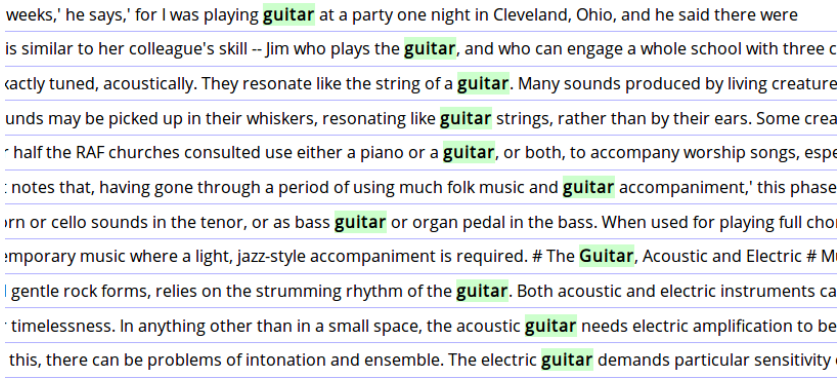
1.5 irudia – Geruza anitzeko pertzeptroia. Iturria `goo.gl/kkCEsd`.

y konbinaketa linealari f aktibazio funtzio batetik pasatzen du[5] eta, atalasebalio baten arabera, irteera bitarra kalkulatzen du (sailkatzaile lineala da). Irteera bitar horren eta benetako irteeraren (f funtzioaren ondorengoa) arteko desberdintasunaren funtzioan, w_i pisuak eguneratzen ditu, eta, hain zuzen, eguneraketa horren bitartez ikasten du. Ohikoena 1.5. irudiko adibidea bezalako arkitekturak dira, elkarrekin lotutako hainbat pertzeptroiz osatutakoak. Aipatutako adibideak, esaterako, bi geruzatako arkitektura erakusten du.

Egungo neurona-sareak neurona anitzez osatuta daude, ikasketa-prozesu eta arkitektura askoz konplexuagoa dituzte, baina, pertzeptroiaren egitura eta ikasketa-prozesua dituzte oinarri. Azken urteotako konputazio-gaitasunaren gorakada dela-eta, beren erabilerak gorakada ikusgarria izan du hainbat ikerketa-ildotan, hizkuntzaren prozesamendua barne.

Gauzak horrela, XXI. mendearen hasieran hizkuntza ereduak neurona-sareen arkitekturan integratzeko aurreneko proposamenak plazaratu ziren (Bengio *et al.*, 2003), eta tesi-lan honetan erabilitako ereduak azken horien bertsio hobetuak eta optimizatuak dira. Bada, hizkuntza ereduak ele jakin bateko hitzen sekuentzien tasun estatistikoak jasotzen dituzte; hau da, aurreko hitzak jakinik, hurrengo hitza aurresaten dute. (1.2) ekuazioak hizkuntza-eredu soil bat deskribatzen du:

[5] Funtzio ez-linealak dira, rektifikatzailea, sigmoidea edo tangente hiperbolikoa.



1.6 irudia – *Guitar* hitzaren agerkidetzak *Brittish National Corpusean*. Iturria `https://corpus.byu.edu/bnc/`.

$$P(w_1, ..., w_m) = \prod_{i}^{m} P(w_i | w_1, ..., w_{i-1}) \tag{1.2}$$

Ekuazio horretan m corpusaren hitz kopurua da eta w_i behatutako hitza. Hizkuntza horren P hizkuntza eredua osatzen aldera, w_i guztien probabilitateak kalkulatzen dira beren aurreko sekuentzien bitartez. Oro har, hitzen probabilitateak kalkulatzeko n-gramak baten baitako sekuentziekin kalkulatzen dira.

Hizkuntza-ereduetan oinarritutako iragarpenak neurona-sareetan integratzeak ikaragarrizko arrakasta izan dute hizkuntzaren prozesamenduan. Hizkuntza-ereduen metodo horretaz baliatzeaz gain, hitzen esanahia beren inguruko distribuzioaren bitartez inferitzen dute (ik. 2.1. atala), hots, testu corpusetako hitz-testuinguru agerkidetzen bitartez. Esaterako, 1.6. irudiak *guitar* hitzaren agerkidetzak erakusten ditu *Brittish National Corpusean*[6]. Adibide horri jarraiki, *guitar* hitzaren esanahia (leiho baten baitako) bere testuinguru guztien esanahien funtzioan kalkulatuko da, hala nola, *rhythm*, *string*, *acoustic*, *electric*, *piano*, *organ* hitzen esanahiekin. Testuinguru horietan ikusten dugunez, hitz-testuinguru agerkidetzetako erlazioak antzekotasunezkoak (*guitar-piano*, *guitar-organ*) zein ahaidetasunezkoa (*guitar-rhythm*, *guitar-string*) izan daitezke.

Hala, hitzen esanahiak bektore-espazio trinko batean kodetzen dituzte, dimentsio anitzetako bektore eskalarretan. Tesi-lan honetan testu-corpusetik

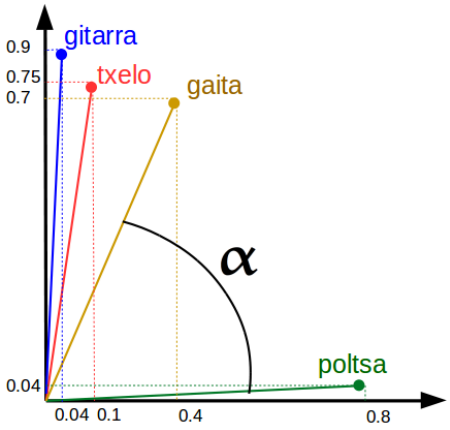
[6] `https://corpus.byu.edu/bnc/`

erauzitako hitzen errepresentazio distribuzional horiei hitz-bektore (*embedding*) izendatuko ditugu. Pertzeptroiaren ikasketan bezala, hitzen esanahiak (hitz-bektoreak) corpusa prozesatu ahala eta iterazioz iterazio eguneratu beharreko pisuak dira, optimizatu beharreko parametroak. Hitz-bektoreen dimentsioak hitzen ezaugarri semantiko latenteak dira.

Hitzen esanahiak bektore-espazio trinko batean izanik, hitzen arteko antzekotasuna kalkulatzea erraza da; 1.3. irudiko adibideko hitzen antzera, hitz-bektore bakoitzak ehunka ezaugarri latente esleituta ditu, eta, antzekotasuna ezaugarrien arteko konparaketa kuantitatibo batekin erreproduzitu daiteke[7]. Esaterako, 1.7. irudiak 1.3. irudiko lau hitzen hitz-bektoreen aurreneko bi dimentsioekin osatutako bektore-espazio bat azaltzen du. Hitz bakoitzaren dimentsioak koordenatu bezala ulertzen direnez, hitz bakoitza bi dimentsiotako espazio batean kokatzen da. Bada, 1.7. irudiarekin lotuta, 1.2. taulak hitzen hitz-bektoreen arteko angeluak (α) eta azken horien kosinuak ($cos(\alpha)$) adierazten ditu, hots, hitzen arteko antzekotasun semantikoa. Adibide honetan ikusten denez, zenbat eta zero gradutik hurbilago egon α, orduan eta antzekotasun handiagoa dute hitz-bektoreek. Ordea, zenbat eta laurogeita hamar gardutik hurbilago egon, orduan eta antzekotasun baxuagoa daukate. Gauzak horrela, *gitarra-txelo* bikoteak oso angelu baxua daukanez ia antzekotasun maximoa kalkulatu dugu, *gaita-poltsa* bikotearentzat antzekotasun-balioa ertaina da, eta *gitarra-poltsa* parean ia ez dago antzekotasunik.

[7]Kosinu-antzekotasunaren bidez kalkulatzea oso zabaldua dago. a eta b hitz-bektoreen arteko antzekotasuna bien arteko α angeluaren kosinua da: $cos(\alpha) = \frac{a \cdot b}{|a| \cdot |b|}$. Emaitza hori [-1,+1] tartean dago, beraz, zenbat eta unitatetik hurbilago egon, orduan eta antzekoagoak hitz-bektoreak.

1.8 irudia – Bi dimentsiotako bektore-espazioa, hainbat kategoria semantikorekin. Kategoria bakoitzeko hitzak kolore berarekin daude, eta espazioan multzokatu egiten dira. Iturria Li *et al.* (2016).



1.7 irudia – Hitz-bektoreak.

Hitz bikotea	α	$cos(\alpha)$
gitarra-gaita	27,12°	0,889
gitarra-txelo	5,05°	0,996
gitarra-poltsa	84,59°	0,094
gaita-txeloa	22,15°	0,926
gaita-poltsa	57,39°	0,539
txeloa-poltsa	79,54°	0,181

1.2 taula – Kosinuaren antzekotasuna.

Gainera, bektore-espazio horien berezitasun nagusietakoa antzeko hitzak multzokatu egingo direla da, antzeko testuinguruetatik ikasiko baitira. 1.8. irudiak, esaterako, bi dimentsiotako bektore-espazio bat irudikatzen du, eta kategoria bereko hitzak multzokatuta agertzen dira. Hitzen arteko antzekotasuna eta hitzen distribuzioak ertsiki lotuta daudela kontua izanik, hitzen arteko antzekotasun-emaitzak hobetzeko beren distribuzioetan eragin beharko dugu. Azken hori da, hain zuzen, hurrengo kapituluetan (metodo desberdinekin) egingo duguna.

Kontuan izan, metodo horiek agerkidetzetan oinarritzen direla, eta azken hori muga serioa dela hitzen errepresentazioak kalkulatzerakoan, hitzen semantikaren hainbat ñabardura ez baitira islatzen agerkidetzetan. Agerkidetzetako informazio semantikoa nahiko azalekoa izaki, testuetan oinarritutako metodoek bi desabantaila dituzte:

- Antzekotasun- eta ahaidetasun-erlazioak nahastu egiten dira hitz-bektoreetan. Azken horiek, oro har, ondo jasoko dituzte antzekotasuna eta ahaidetasuna, baina, ez dira onak izango batean ez bestean.

- Agerkidetzetan ez dago hitzen arteko inongo erlazio espliziturik, erlazioak latenteak dira. Ondorioz, hitz-bektoreak ikasterakoan ezin da hitzen erlazioen inongo kontrolik burutu.

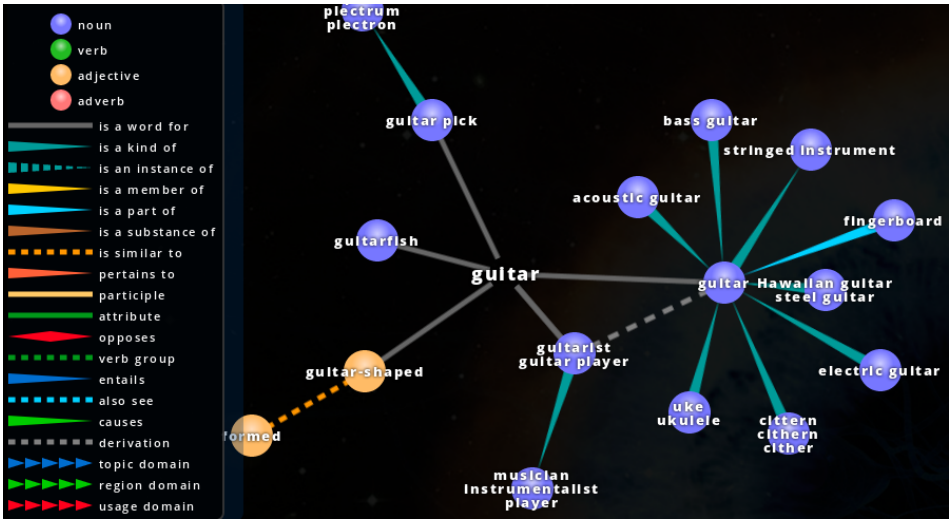
Corpusetan oinarritutako metodoen muga horiek gainditze aldera, tesi-lan honetan ezagutza-baseetan oinarritutako metodoez ere baliatu gara. Hurrengo atalean azken horien deskribapen laburra egingo dugu.

Ezagutza-baseetan oinarritutako metodoak

Metodo hauetan testu-corpusen oso bestelako informazio semantikoa ustiatzen da. Testu agerkidetzetan ez bezala, ezagutza-baseetan informazioa esplizitua da, eta, ondorioz, semantikaren hainbat ñabardura jasotzeko gai. Tesi-lan honi dagokionez, azken ezaugarri hori da interesatzen zaiguna, hitzen errepresentazioetan antzekotasuna edota ahaidetasuna modu kontrolatuan kodetzea ahalbidetuko baitigute.

Hizkuntzaren prozesamenduan ezagutza-base lexikalak ohikoak dira. Azken horiek ezagutza linguistikoa modu egituratuan gordetzen dute, ele bateko lexikalizazioak beren arteko erlazioekin lotuz. Ezagutza-base lexikalen artean WordNet (Miller, 1995) aspaldidanik erabili izan da hizkuntzaren prozesamenduan. WordNet jatorrian ingeleserako sortua izan zen, eta bere ezaugarri nagusiak hurrengoak dira: hitzak *synset* izeneko sinonimo-multzoetan batzen ditu, *synsetak* erlazio semantikoen bidez lotuz, eta hitzen definizioak eta adibideak ere eskainiz. Bada, 1.9. irudiak WordNeten *guitar* hitzaren auzokideak eta beren arteko erlazioak erakusten ditu. Adibide horretan *guitar* lexikalizazioaren adiera[8] usuenari so, musika-tresnaren adierari, bere

[8]Bada, *guitar pick*, *guitarfish*, *guitar-shaped*, *guitarrist* eta *guitar*. Adieren disanbiguazioa ezagutza-baseekin lotutako ataza bada ere, tesi-lan honetan ez dugu ildo hori jorratu.



1.9 irudia – WordNet ezagutza-basean *guitar* hitzaren testuinguruko hitzak, beren erlazio semantikoekin batera. Iturria `https://visuwords.com`.

auzokideak gehienbat antzeko hitzak dira (*bass guitar*, *stringed instrument*, *Hawaiian guitar*), baina, hitz ahaideren bat ere agertzen da (*fingerboard*). Oro har, ezagutza-base gehienetan antzekotasun-erlazioak gailentzen dira.

Adibide horretan lexikalizazioen arteko loturak erakutsi baditugu ere, ezagutza-baseetan loturak kontzeptuen artean egiten dira, eta kontzeptuei lexikalizazioak esleitzen zaizkie. Kontzeptu horiek hizkuntzaren independenteak dira, eta, ondorioz, sarritan ezagutza-baseak eleartekoak izaten dira (Camacho-Collados *et al.*, 2015; Agirre *et al.*, 2012), ele desberdinek azpian kontzeptuez osatutako egitura bera partekatu baitezakete. Ezaugarri hori ere ustiatuko dugu gure tesi-lan honetan, elearteko hitzen antzekotasun- edota ahaidetasun-emaitzak hobetzeko balioko baitigu.

Gure ikerketa-ildoan ezagutza-baseen gainean algoritmo globalak deiturikoak aplikatu ditugu (Agirre *et al.*, 2009b). Azken horiek ezagutza-basea grafo baten moduan ulertzen dute, eta bere informazio estruktural osoa prozesatzen. Horretarako, algoritmoek ezagutza-baseko kontzeptuak erpin bezala ulertzen dituzte, eta azken horien arteko erlazioak ertz bezala. Tesi-lan honetan WordNet (Miller, 1995) ezagutza-basean oinarritu gara, hizkuntzaren prozesamenduan oso erabilia baita hainbat atazatan (Leacock and Chodorow, 1998; Banerjee and Pedersen, 2002; Budanitsky and Hirst, 2006; Agirre

et al., 2009a), hitzen arteko antzekotasuna barne.

1.3 Aurrekariak IXA taldean

Tesi-lan hau IXA taldean[9] burutu da, Euskal Herriko Unibertsitatearen baitan dagoen ikerketa-taldean. IXA taldea hizkuntzaren prozesamenduaren alorrean dabil, eta bere ikerketa euskararen inguruan ardazten badu ere, beste hizkuntzekin ere lan egiten du. Tesi-lan hau IXA-taldean kokatze aldera, lotutako lan batzuk aipatuko ditugu.

Ezagutza-baseen inguruko lanari dagokionez, IXA taldean euskarazko WordNeta (Pociello *et al.*, 2011) osatu zen 2000. eta 2003. urteetako EuroWordNeten diseinuei jarraiki, eta, ondoren, euskarazko WordNeta *Multiligual Central Repositoryra* (Agirre *et al.*, 2012) mugitu zen. Gainera, IXA taldean ezagutza-baseen egitura semantikoak ustiatzeko UKB tresna garatu zen (Agirre *et al.*, 2009b), azken horiekin kalkulatutako bektoreekin antzekotasun-atazan lan eginik (Agirre *et al.*, 2010).

Tesi-lan honen aurrekari gisa, IXA taldean *Semantic Textual Similarity* eta *Typed Similarity* atazak ere landu dira (González Aguirre, 2017) . Aurrenekoa esaldien antzekotasuna neurtzeko ataza da, eta, bigarrenak, hiztegi digital bateko antzeko entitateen arteko erlazioak identifikatzean datza. Horiez gain, Lopez-Gazpio *et al.*-ek (2017) *Interpretable Semantic Textual Similarity* ataza ere landu du, eta aurretik aipatutako *Semantic Textual Similarity* atazan interpretagarritasun geruza bat txertatu.

1.4 Motibazioa eta ekarpenak

Tesi-lan honen motibazioak hurrengo puntuetan laburbiltzen dira:

- Hitzen errepresentazioen arteko antzekotasun-emaitzak hobetzea: antzekotasuna hizkuntzaren prozesamenduko hainbat atazen oinarrian dago, eta hitzen errepresentazioen kalitatearen adierazgarria da.
- Hitzen errepresentazio hobeak lortzeko metodoak eta baliabideak garatzea: antzekotasuna eta hitzen esanahia ertsiki lotuta daude, eta ezin bata bestea barik ulertu. Hala, antzekotasuna hobetu nahi badugu errepresentazio hobeak beharko ditugu.

[9] `http://ixa.si.ehu.es/`

- Aurreko biak elearteko bektore-espazioetan inplementatzea.

Aipatu bezala, antzekotasunaren inguruko ikerketa-ildoa gehienbat corpusetan oinarritutako metodoen eta ezagutza-baseetan oinarritutakoen hitzen errepresentazioekin (ik. 1.2. eta 1.2. atalak) egin da. Tesi-lan honen hastapenetan ez zegoen ia bi iturri horiek konbinatzeko proposamenik, eta, bat bera ere ez espazio eleaniztunetan egiten zuenik. Hala, aurreko paragrafoko hiru puntuetan aipatutakoak gauzatze aldera, hipotesi nagusi batean eta azken horretatik eratorritako hiru azpi-hipotesitan ardaztu dugu gure ikerketa-ildoa:

- Testu-corpusetako eta ezagutza-baseetako informazio semantikoa desberdina baina osagarria da.
  - Bi baliabide horiek errepresentazio hibrido batean konbinatuz gero, errepresentazio hobeak lortuko dugu.
  - Errepresentazio hibridoek hitzen arteko antzekotasun- eta ahaidetasun-emaitzak hobetuko dituzte.
  - Esandako guztiak elearteko espazioetan ere betetzen dira.

Hipotesi horiek guztiek gure ikerketa-ildo osoa ardaztu dute. Lehenik, gure hipotesiak errepresentazio elebakarretan gauzatu ditugu, eta, ondoren, eleartekoetara hedatu. Hurrengo puntuek lan honetako ekarpen nagusiak laburbiltzen dituzte, dagozkien kapituluekin batera:

- Ausazko ibilbideen algoritmo batez baliatuz, ezagutza-base bateko informazio estrukturala inplizituki corpus bateko agerkidetzetan kodetu dugu (3. kapitulua).
- Ezagutza-baseetan oinarritutako corpusetatik abiatuta, azken horren hitzen errepresentazio distribuzionalak kalkulatu ditugu (3. kapitulua).
- Ezagutza-baseak eta testu-corpusak konbinatzen dituzten errepresentazio distribuzional hibridoak sortzeko metodo eraginkorrak proposatu ditugu (3. eta 4. kapituluak).
- Aurreko guztien bertsio elebidunak 5. kapituluan landu ditugu.

Ekarpen horiek guztiak 6. kapituluan sakonago aztertuko ditugu .

1.5 Tesi-lan honetatik ateratako argitalpenak

Tesi-lan hau hiru argitalpenetan oinarritzen da, eta azken horiek kapituluka sailkatu ditugu hurrengo zerrendan:

- 3. kapitulua - Ezagutza baseetan oinarritutako teknikak: ausazko ibilbideak:
  - Goikoetxea J., Soroa A., Agirre E., and Donostia B.C. **Random walks and neural network language models on knowledge bases**. Proceedings of HLT-NAACL, pp. 1434-1439, ISBN-978-1-937284-73-2, 2015.
- 4. kapitulua - Testu eta ezagutza-baseen konbinaketa:
  - Goikoetxea J., Agirre E., and Soroa A. **Single or multiple? combining word representations independently learned from text and wordnet**. Proceedings of AAAI, pp. 2608-2614, ISBN-978-1-57735-760-5, 2016.
- 5. kapitulua - Eleaniztasunera hedapena:
  - Goikoetxea J., Agirre E., and Soroa A. **Bilingual embeddings with random walks over multilingual wordnets**. Knowledge-Based Systems, 2018.

Hurrengo argitalpenak ere, tesi-lan honekin zerikusia daukate:

- Goikoetxea J., Agirre E., and Soroa A. **Exploring the use of word embeddings and random walks on Wikipedia for the CogAlex shared task**. Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex), 31-34, ISBN-9781634392174, 2014.
- Goikoetxea J., Agirre E., Soroa A.**Konbinatu eta Irabazi! Hitzen Semantikaren Errepresentazio Osoagoaren Bila**. Ikergazte, 2015.
- Goikoetxea J., Lopez-Gazpio I., Agirre E., Maritxalar M., Soroa A.**Testuloturen labirinto semantikoan barna, esanahi-bektoreak lagun!**. Ikergazte, 2017.

1.6 Tesi-txostenaren egitura

Tesi-lan hau hurrengo kapituluetan egituratuta dago:

- 2. kapitulua - Aurrekariak:

  Aurrekariak hurrengo azpi-ataletan banatu ditugu: aurrenekoan eta bigarrenean Semantika Distribuzionala eta azken horretatik eratorritako eredu distribuzionalak deskribatuko ditugu; hirugarrenean ezagutza-baseetan oinarritutako metodoen nondik norakoak; laugarrenean aurreko bi familiak konbinatzen dituzten metodoak; bosgarrenean ebaluaziorako antzekotasun urre-patroi elebakarren eta eleartekoen ezaugarriak.

- 3. kapitulua - Ezagutza baseetan oinarritutako teknikak: ausazko ibilbideak:

  Tesi-lan honen muinean dauden oinarrizko baliabideak sortzeko esperimentuak deskribatuko ditugu. Lehenik, ezagutza-baseetatik corpusak eraazteko ausazko ibilbideetan oinarritutako metodoa aurkeztuko dugu, eta, ondoren, corpus mota berri horietatik errepresentazio trinkoak[10] kalkulatzeko neurona-sarearen ikasketa-prozesua aztertuko dugu. Gero, ezagutza-base batean oinarritutako corpusak eta errepresentazio trinkoak lortzeko esperimentuak deskribatuko ditugu, erdietsitako emaitzekin batera. Atal hori esperimentu horietatik ateratako ondorioekin bukatuko dugu.

- 4. kapitulua - Testu eta ezagutza-baseen konbinaketa:

  Testu eta WordNet espazio bereizietako informazio semantikoa konbinatzeko hainbat metodo aurkezten hasiko gara. Esperimentuen atalean, konbinaketak hainbat antzekotasun urre-patroitan ebaluatuko ditugu, eta artearen egoerako metodo baliokideekin konparatuko. Esperimentuen azkenengo atalean, bi baliabide semantiko baino gehiago konbinatuko ditugu, eta, azken horrek antzekotasunean duen eragina aztertuko dugu. Esperimentuekin bukatu ostean, konbinaketekin ondorioztatutakoak laburbilduko ditugu.

[10] Tesi-lan honetan ezagutza-baseetako hitzen errepresentazio distribuzionalei ez diegu hitz-bektore deituko, errepresentazio trinko baizik.

- 5. kapitulua - Eleaniztasunera hedapena:

  Aurreko bi kapituluetan proposatutako metodoak eta errepresentazio hibridoak elearteko espazioetara hedatuko ditugu. Lehenik, 1.2. atalean deskribatutako neurona-sare bati ezagutza-baseetatik eratorritako murriztapenak txertatzeko gure proposamena deskribatuko dugu. Ondoren, ezagutza-baseetan oinarritutako ausazko ibilbide elebidunen metodoa azalduko dugu, eta baita azken horretatik osatutako corpus elebidunen ezaugarrienak ere (bai metodoa eta bai corpusa 3. kapituluan proposatutakoaren hedapen elebidunak dira). Esperimentazioaren atalean, erabilitako baliabideak laburtu eta gero, gure errepresentazio elebidunak artearen egoerako bi metodorekin konparatuko ditugu elearteko antzekotasunean, hainbat hizkuntza-bikotetan.

- 6. kapitulua - Ondorioak eta etorkizuneko lanak:

  Tesi-lan honen sintesi moduan, gure ikerketa-ildoari eginiko ekarpenak zerrendatuko ditugu, eta tesian ateratako ondorio nagusiak laburbilduko. Azkenik, etorkizunean egiteko lanekin bukatuko dugu.

2

Aurrekariak

Atal hau gure ikerketa-ildoko arloaren egoeraren sintesia da. Sarreran aipatu bezala, tesi-lan honen muinak ertsiki erlazionatutako bi aurpegi ditu, hots, hitzen arteko antzekotasuna eta hitzen errepresentazio semantikoak. Alor horiek hizkuntzalaritzan, psikologian eta filosofian sakon errotuta daude, eta aurreneko ikerketak eredu konputazionalen sorrera baino hamarkada batzuk aurretik burutu ziren. Tradizio horietako aurrekariak izango da, hain zuzen, atal honen abiapuntua, gure eredu konputazionalen oinarri teorikoak baitira. Oinarri teorikoak azaldu ostean, lan honetan erabilitako ereduen familiak eta euren nondik norakoak deskribatuko ditugu. Azkenik, gure esperimentuetako hitzen errepresentazio semantikoak ebaluatzeko baliabideak eta irizpideak deskribatuko ditugu.

2.1 Semantika Distribuzionala

Sarrerako atalean antzekotasuna gizakiok errealitatea egituratzeko eta ordenatzeko tresna kognitiboa garrantzitsua dela ikusi dugu, eta, ondorioz, makinek errealitatea ulertzea gura badugu gaitasun hori erreproduzitu beharko dute. Hizkuntzaren prozesamenduari dagokionez, antzekotasuna erreproduzitzeko gakoa hitzen tasun semantikoen errepresentazioak direla nabarmendu dugu, antzekotasuna eta esanahia ertsiki lotuta baitaude. Antzekotasunak hizkuntza naturalean duen funtzioa[1] oinarri legez hartuta, hizkuntzalaritzaren tradiziotik hitzen esanahiak modelatzeko Semantika Distribuzionala pro-

[1] Hots, errealitatea egituratzea eta ordenatzea.

posatu zen, egun hizkuntzaren prozesamenduan oso erabilia. Bada, Semantika Distribuzionala Harris-en (1954) hipotesi distribuzionalean oinarritzen da, eta bere muina hurrengo esaldian laburbildu daiteke:

Esanahi antzeko hitzek testuinguru berean agertzeko joera dute.(Harris, 1954)

Esaldi horrek, bere laburtasunean, hitzen esanahien eta antzekotasunaren arteko loturaren inguruko intuizio soil bezain sakona (eta praktikoa) erakusten digu; hitzen esanahiaren ikuspegitik, esanahia modelatzeko bere testuinguruko informazioaren bitartez egin beharko da; antzekotasunaren ikuspegitik, testuinguruen informazioak hitzen esanahia definitzen duenez, hitzen arteko antzekotasuna beren testuinguruen menpekoa da.

Hipotesi horietaz baliatuz, hizkuntzaren prozesamenduan eredu semantiko distribuzionalak (ESD) inplementatu dira eta oso hedatuta daude, hainbat atazatan arrakastatsuak izan baitira. Hizkuntzaren prozesamenduan, baina, "esan gabe doa" jarrera daukagu Harris-en (1954) hipotesiaren inguruan, ezer gutxi dakigu bere jatorriaz. Zeintzuk dira, baina, hipotesi distribuzionalaren erroak?

2.1.1 Semantika Distribuzionalaren historiaurrea

Esan bezala, Semantika Distribuzionalaren XX. mendearen erdialdean dauka jatorria, hipotesi distribuzionalean. Hala ere, hizkuntzalaritzaren tradizioan errotutako hipotesi horren "arbaso teorikoak" XX. mendearen hasierakoak dira; hain zuzen, Ferdinand de Saussuren eta Wittgensteinen teoria linguistikoak.

Saussurren teoria linguistikoen itzala hizkuntzalaritzaren esparrutik haratago heldu zen, filosofian, antropologian eta soziologian ere oihartzuna izan baitzuen[2]. Bada, hizkuntzalari suitzarrak hizkuntza zeinuez osaturiko egitura legez ulertzen zuen, eta, zeinuei barik, egitura horren azpiko legeei ematen zien garrantzia. Bere aburuz, zeinuak beren desberdintasun funtzionalen bitartez identifikatzen dira, eta funtzio horiek zeinuek hizkuntzaren sistemaren baitan duten rola definitzen dute. Kontuan izan, diferentzia funtzional horiek sistema baten baitan dutela zentzua soilik, isolatutako zeinuak ezin baitu desberdintasun-erlaziorik izan. Ikuspuntu horretatik, sistema (esaterako, hizkuntza) desberdintasun funtzionalen elkarrekintza da:

[2]Hizkuntzalari suitzarra XX. mendean eskola filosofiko garrantzitsuenetakoen aitatzat hartzen da, estrukturalismoa.

Hizkuntzan, desberdintasunak baino ez daude. (Saussure, 1916/1983)

Saussurren sistemen teoriak hipotesi distribuzionalaren inspiratu zuela kontuan izanik, ez da zaila bien arteko paralelismoak aurkitzea: hizkuntza sistema baten legez ulertuz gero, hitzak zeinuak lirateke, hitzen arteko erlazio semantikoak zeinuen arteko elkarrekintzak eta hitzen esanahiak zeinuen desberdintasun funtzionalak. Esan gabe doa, ikuspegi horretatik desberdintasun funtzionalen gradua hitzen antzekotasunarekin parekatu dezakegula.

Saussurez gain, Ludwig Wittgenstein filosofo austriarra ere hipotesi distribuzionalaren aitzindaritzat hartzen da. Saussurek hizkuntzaren egituran jarri bazuen arreta, Wittgensteinek logikaren izaeran eta "berbekin esan eta uler daitekeenaren mugetan"[3] egin zuen. *Philosophische Untersuchungen*[4] hilondoko lanean zera esan zuen:

Hitz baten esanahia hizkuntzan duen erabilera da. (Wittgenstein, 1953)

Wittgensteinek esaldi hori hurrengo moduan birformulatu genezake: hitz baten esanahia hizkuntzan dituen testuinguruez osatutako informazioa da. Saussurren antzera, Wittgensteinek hitzen esanahia ez du modu isolatuan ulertzen, hizkuntzaren sarearen baitan kokatzen baitu, baina, egiturari barik, erabilerari ematen dio garrantzia.

Aipatutako bi autoreek eragin handia izan zuten XX. mendeko hizkuntzalaritzan, eta, beren ikuspegi horietan oinarrituta, hainbat hizkuntzalarik erabilpenera bideratutako metodologiak garatu zituzten hizkuntzaren ikerketan. Testuinguru intelektual horretan sortu zen, hain zuzen, atal honen hasieran azaldutako hipotesi distribuzionala, besteak beste. Harrisek hasieran analisi fonemikoaren baitan aurkeztu bazuen ere (Harris, 1951), gero maila linguistiko orori aplikatzeko metodologia izatera pasatu zen (Harris, 1954, 1970).

Hala, hipotesi distribuzionalaren garai eta ildo berean, itzulpenaren inguruko bere lanean eta bere garaiari asko aurreratuz, Weaver-ek (1955) itzulpen automatikoan hitzen desanbiguazioa hitz-xedearen inguruko auzokideen frekuentzietan oinarrituta egon beharko lukeela esan zuen. Harrisek eta Weaverrek esandakoaren harira, hona hemen Firthen hitzak:

[3]Bere aforismo ezagunenetakoa *Tractatus Logico-Philosophicus* (Wittgenstein, 1921/2013) lanekoa da, eta zera dio: "Berba egin ezin denaren inguruan, isilik mantendu behar da".

[4]Euskaraz "Ikerketa filosofikoak".

Hitz baten esanahia haren auzokideen bidez ezagutuko duzu. (Firth, 1957)

Berba desberdinak erabiltzen badituzte ere, atal honetan aipatutako autoreek gauza bera nabarmentzen dute; hots, *hitz baten esanahia bere testuinguruan datza*. Esan bezala, bai Harrisen, Firthen eta Weaverren ikuspegia enpirizista zen, ikerketa linguistikorako metodologia bilatu guran baitzebiltzan. Ikuspegi horrek, noski, hitzen esanahiak Semantika Distribuzionalaren bidez eredu konputazional batean modelatzea ahalbidetu zuen, gaur egungo ESDen sorkuntzarako bidea urratuz.

2.1.2 Errepresentazio distribuzionalak, ESD

Aurreko atalean azaldutako hipotesi distribuzionala hiru hamarkada baino gehiagotan hizkuntzalaritzaren eremura soilik mugatu zen. Hipotesi horretan oinarrituta, 80. hamarkadaren amaieran eta 90. hasieran Deerwester *et al.*-ek (1990) Harrisen (eta beste autoreen) esanahiari buruzko intuizioak algoritmo praktiko batean gauzatu zituen; hain zuzen, *Latent Semantic Analysisen* (LSA). Azken hori da aurreneko ESDa eta hitz-dokumentu kontaketa-matrize batean oinarrituta dago. Laster agertu ziren LSAren bertsio finduagoak ere (Blei *et al.*, 2003; Lund, 1995), eta egun oraindik erabiltzen dira. Metodo horiek guztiak etiketatu gabeko corpusak prozesatzen dituzte, modu ez gainbegiratuan, eta hitzen esanahia bektore-espazio batean kodetzen dute. Kontuan izan, hitzen esanahia modelatzeko (eta, ondorioz, antzekotasuna erreproduzitzeko) oinarria hitz-testuinguru agerkidetzetako informazioa dela.

XXI. mendearen hasieran, hizkuntza ereduak neurona-sareetan integratu eta kontaketa-metodoetatik iragarpen-metodoetara (Bengio *et al.*, 2003; Collobert and Weston, 2008; Turian *et al.*, 2010) egiten da jauzi. Corpus osoko hitz-testuinguru kontaketak matrize batean jaso barik, leiho baten baitako hitz-sekuentzietako agerkidetzen bidez kalkulatzen dira esanahiak. Urte batzuk beranduago, konputazio-gaitasunaren gorakadaren eskutik, neurona-sareetan oinarritutako metodo horiek optimizatu eta hobetu egin dira (Mikolov *et al.*, 2013a; Pennington *et al.*, 2014), hizkuntzaren prozesamenduan arreta ikaragarria erakarriz. Inplizituki bada ere, ESD belaunaldi berri hau hitz-testuinguru matrizeez baliatzen da (Levy and Goldberg, 2014), eta, ondorioz, hitzen esanahia eta antzekotasuna modelatzeko kontaketa-metodoen antz-antzeko estrategia darabil.

Hurrengo bi ataletan aurreko paragrafoetan aipatutako ESD moten ezaugarriak deskribatuko ditugu: kontaketetan eta iragarpenetan oinarritutako metodoak.

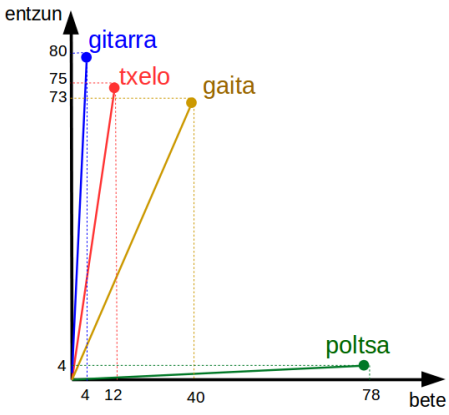
2.1.3 Kontaketa-metodoak

Atal honen sarreran aipatu bezala, Deerwester *et al.*-ek (1990) hitzen esanahia modelatzeko aurreneko ESDa proposatu zuten, LSA. Gerora, LSAren bertsio finduagoak ere agertu ziren, hala nola, *Probabilistic Latent Semantic Analysis* (Hofmann, 1999), *Hyperspace Analog to Language* (Lund, 1995; Lund and Burgess, 1996) eta *Latent Dirichlet Allocation* (Blei *et al.*, 2003) ereduak.

Oro har, kontaketa-metodo horiek guztiek ezaugarri berberak dituzte. Bereizgarri nagusia esanahia bektoreen bidez adierazten dituztela da, eta azken horiek testu corpusetako hitzen agerkidetzetako informazioaz osatuta daudela. Bada, bektore horiek agerkidetza-matrize bateko lerroak dira; lerro bakoitza xede-hitz bati dagokio, eta zutabe bakoitza xede-hitz horrek testuinguru jakin batean izandako agerkidetzaren funtzioa izango da. Testuinguruak modu desberdinetan definitzen dira metodoaren arabera, sinpleetatik (dokumentu edo hitz agerkidetzak, esaterako) konplexuagoetara doaz (erlazio sintaktikoekin erlazionatutako hitzak, esaterako).

	entzun	bete	lan	...
...	...	...	...	...
gitarra	80	4	7	...
txeloa	75	12	5	...
gaita	73	40	9	...
poltsa	4	78	46	...
hartz	30	15	8	...
eguzki	7	31	29	...
...	...	...	...	...

2.1 taula – Agerkidetza-matrizearen adibidea.



2.1 irudia – ESD adibidea.

2.1. taulan testu-corpus batetik erauzitako agerkidetza-matrize baten kontaketak agertzen dira. Bada, 1.3. irudiko lau terminoen tasunak agerkidetzen kontaketak bihurtu dira, ESD bateko errepresentazio bilakatuz. 2.1.

irudiak agerkidetza-matrize horren *entzun* eta *bete* dimentsioez osatutako bektore-espazioa aipatutako lau hitzentzat irudikatzen du. Hipotesi distribuzionalari jarraiki, antzeko kontaketadun hitzak multzokatu egiten dira espazioan, hau da, antzeko esanahiarekin errepresentatzen dira. Adibideko ESDan, *gitarra*, *txelo* eta *gaita* hitzek *entzun* distribuzio antzekoak dituzte[5] eta espazio horretan bata bestetik hurbil daude, hau da, antzekoak dira. *Poltsa* hitzak, ordea, oso bestelako distribuzioa dauka[6], eta nahiko urruntzen da gainontzeko hitzetatik. Hala ere, *gaita* eta *poltsa* hitzak zertxobait hurbilago daude *bete* tasuneko balioa dela-eta.

Hitz-testuinguru kontaketak, baina, ez dira bere horretan uzten, haztapenen bat aplikatzen baitzaie. Hala ere, agerkidetza-matrizearen tamaina handiak izaten dituzte, eta hitzei dagozkien agerkidetzen balioak oso sakabanatuta egoten dira (dimentsio gehienek zero balio dute). Gauzak horrela, konputazioan efizientzia irabazte aldera, dimentsio-murrizketa aplikatzen zaie agerkidetza matrizeei. Kontuan izan, murrizketa aplikatu osteko dimentsioak ezin direla zuzenean behatu, latenteak baitira. Izan ere, murriztapenaren aurretik dimentsioek zuzenean agerkidetzei egiten diete erreferentzia; murriztapenaren ondoren, dimentsioek agerkidetzak modu inplizituan jasotzen dituzte. Beste berba batzuetan esanda, hitzen esanahiak dimentsio konkretuen bidez osatuta egotetik, dimentsio abstraktuen bidez osatuta egotera pasatzen dira. Esaterako, aurreko adibideko *gitarra* hitzaren esanahiak milaka dimentsio izango lituzke agerkidetza-matrize batean (corpuseko hiztegiaren sarreren beste), eta berarekin agertutako testuinguru bakoitzerako kontaketen informazioa eskuragarri legoke. Esaterako, *entzun* hitzarekin 80 bider agertu dela, *bete* hitzarekin 4 eta *lan* hitzarekin 7. Azken horiei haztapenen bat aplikatuta ere, agerkidetzen informazioa zuzenean behatu daiteke. Murriztapenaren ondoren matrizearen dimentsionaltasuna nabarmen jaitsiko da (ehunka batzuk izaten da ohikoena), eta dimentsioei eskalarrak esleitzen zaizkie. Hala, *gitarra* bektoreak hurrengo itxura izango luke bost dimentsiotan:

gitarra 0.0211 -0.0234 0.0392 0.0001 -0.1064

Murriztapenaren ondoren, bada, bektore-espazio trinko bat dugu, eta azken horrek horrek jatorrizko agerkidetza-matrizeko informazioa gordetzen du,

[5]Hirurek *entzun* dimentsioan balio altua dute, eta *bete* dimentsioan balio baxuak edo, gaitaren kasuan, ertaina.

[6]*Entzun* dimentsioan balio baxua, eta *bete* dimentsioan altua.

modu latentean. Bektore-espazio horretako lerroak hitzen errepresentazio distribuzional trinkoak dira, hots, hitz-bektoreak.

Gauzak horrela, kontaketa ESDek iragarpen-metodoen oinarriak ezarri zituzten. Hurrengo atalean neurona-sareetan oinarritutako iragarpen-metodoak azalduko ditugu, azkenengo urteotan hizkuntzaren prozesamenduan ikaragarrizko arrakasta izan baitute.

2.1.4 Iragarpen-metodoak

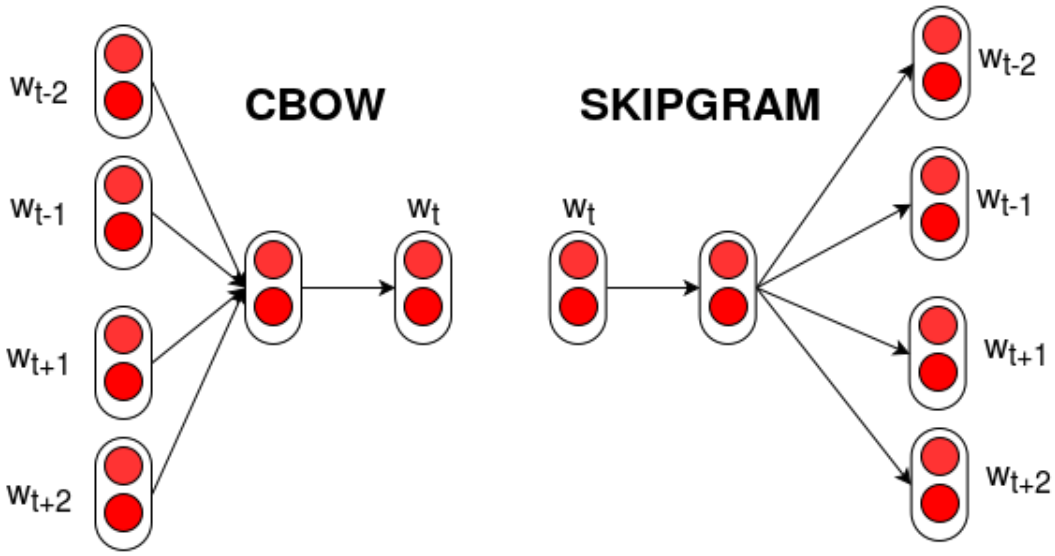
Aurreko atalean deskribatutako kontaketa-metodoetatik abiatuta, azken urteotan ESDen beste belaunaldi bat jalgi da. Labur esanda, ESD berrietan hitz-bektoreak beren testuinguruak aurresateko probabilitatea maximizatzeko helburuarekin kalkulatzen dira (Bengio *et al.*, 2003; Collobert and Weston, 2008; Collobert *et al.*, 2011; Turian *et al.*, 2010; Mikolov *et al.*, 2013a), betiere testu-corpuseko hitz-testuinguru agerkidetzetan oinarrituta. Beraz, kontaketa-metodoetan ez bezala, iragarpen-metodoen ikasketa gainbegiratua da; hots, corpus anotaturik behar ez badu ere, corpuseko hitzen sekuentzietaz baliatzen da iragarpenak egiteko. Azken horiek arrakasta izugarria izan dute hizkuntzaren prozesamenduan, beren dimentsionaltasun trinkoa delaeta oso azkarrak direlako, eta artearen egoera lortu dutelako hizkuntzaren prozesamenduko hainbat atazatan, bai semantikan eta bai sintaxian (Collobert *et al.*, 2011; Socher *et al.*, 2011; Mikolov *et al.*, 2013a; Pennington *et al.*, 2014).

Kontaketa-metodoekin alderatuz gero, iragarpen-metodoek hitz-bektoreen pisuak zuzenean aldatzen dituzte[7], azken horiei dagozkien hitz-testuinguru agerkidetzak optimoki aurresateko. Esan gabe doa, hipotesi distribuzionalari jarraiki, iragarpen ereduetan ere antzeko hitzei antzeko bektoreak esleituko zaizkiela. ESD belaunaldi berriaren aitzindaria Bengio *et al.* (2003) dugu, eta hizkuntza ereduetan oinarritutako neurona-sare bat proposatzen dute. Azken horren eredutik abiatuta, Collobert and Weston-ek (2008) neurona--sare arkitektura orokorragoa proposatzen dute, eta egungo hainbat ereduren oinarria bihurtu da.

Hala ere, hizkuntzaren prozesamenduan hitz-bektoreen erabilerak izandako gorakada Mikolov *et al.*-en (2013a) `word2vec`[8] eredu-multzoari zor zaio. Azken horrek bi eredu ditu, Skip-gram eta Continuous Bag Of Words (CBOW)

[7] Kontuan izan, kontaketa-metodoek lehenik agerkidetza-matrizea sortzen dutela, gero, azken horien pisuak aldatzen, eta, azkenik, dimentsio murrizketa egiten dutela.

[8] `https://code.google.com/archive/p/word2vec/`



2.2 irudia – CBOW eta Skip-gram ereduen arkitekturak. Iturria `goo.gl/nCy3c5`.

izenekoak, hurrenez hurren. Skip-gramen xedea behatutako hitzaren testuinguruko hitzak aurresatea da, eta, CBOWrena, berriz, behatutako hitza aurresatea testuinguruko hitzez baliatuta. 2.2. irudian Skip-gram eta CBOW ereduen arkitekturak deskribatzen ditu. Irudi horretan, CBOW ereduaren sarrerak $w(t-2)$, $w(t-1)$, $w(t+1)$ eta $w(t+2)$ testuinguruko hitzak dira, eta azken horiek geruza ezkutuan bateratu ondoren[9], $w(t)$ behatutako hitza aurresaten dute. Skip-gram ereduan alderantzizkoa egiten da, hots, $w(t)$ behatutako hitza neurona-sareko sarrera geruza ezkutuan sartzen da ,eta azken hori testuinguruko hitzak aurresateko erabiltzen. Aipatutako iragarpen horiek leiho baten baitako hitzekin egiten dira, hau da, erdian dagoen hitza behatutakoa da, eta bere ingurukoak testuinguru hitzak. Zehazki, `word2vec` eredu-multzoko galera-funtzioak leiho baten baitan behatutako hitz-testuinguru agerkidetzen probabilitatea saritzen du, eta ausazko hitz-testuinguruena zigortzen[10].

Iragarpen-metodoen artean agertutako beste eredu esanguratsu bat `GloVe` (Pennington *et al.*, 2014) da, kontaketa-metodoak eta iragarpen-metodoak

[9]Neurona-sareetan geruza ezkutu (*hidden layer*) moduan ezagutzen dira, eta neurona-sarearen irteera bere sarreraren funtzioan kalkulatzeko erabiltzen.

[10]3.1.2. atalean Skip-gram ereduaren galera-funtzioaren xehetasunak azaltzen ditugu.

Probabilitatea eta ratioa	*w=solido*	*w=gas*	*w=ur*	*w=moda*
p(w/izotz)	$1,9 \times 10^{-4}$	$6,6 \times 10^{-5}$	$3,0 \times 10^{-3}$	$1,7 \times 10^{-5}$
p(w/lurrun)	$2,2 \times 10^{-5}$	$7,5 \times 10^{-4}$	$2,2 \times 10^{-3}$	$1,8 \times 10^{-5}$
p(w/izotz)/p(w/lurrun)	8,9	$8,5 \times 10^{-2}$	1,36	0,96

2.2 taula – `GloVe` ereduko ikasketa-prozesua agerkidetza probabilitateekin

konbinatzen baititu ikasketan. Bere egileen aburuz, `word2vec` bezalako metodoek leihoetako informazio lokala soilik prozesatzen dute, eta, hortaz, testu-corpuseko informazio estatistiko globalik ez dute jasotzen. Gauzak horrela, `GloVe` metodoaren galera-funtzioak leiho baten baitako agerkidetzetako informazio lokala eta hitz-testuinguru agerkidetza-matrizeetako estatistika globalak batzen ditu.

`GloVe` ereduaren autoreek hitz-testuinguru esanahia kodetzeko agerkidetzen probabilitate-ratioez (eta ez agerkidetzen probabilitate gordinek) baliatzen dira. Izan ere, ratioek hitz esanguratsuen eta garrantzirik gabekoen artean bereiztea ahalbidetzen dute, eta, ondorioz, probabilitate gordinak baino eraginkorragoak dira semantika kodetzeko. 2.2. taulak agerkidetzen probabilitate gordinekin eta azken horien ratioekin erdietsitako emaitzak azaltzen ditu. Taulako lehenengo bi lerroetan hitz-testuinguru probabilitate gordinak agertzen dira, eta hitza adierazteko *w* erabiltzen da. Esaterako, *izotz* eta *lurrun* testuinguru legez hartuta, 2.2. taulan azken horiekin erlazionatutako *solido* eta *gas* hitzek ratio altua eta baxua dituzte, hurrenez hurren; hots, beren esangura nabarmentzen da ratioen muturreko balioen bidez. Ordea, *ur* hitza *izotz* eta *lurrun* testuinguru biekin hein berean erlazionatuta dago, eta ratioa 1etik hurbil dago. *Moda* hitzaren emaitza ere 1etik hurbil dago, ez baitago batekin ez bestearekin erlazionatuta.

Badira metodo berri anitz, eta beren artean aipagarria da `fastText` (Bojanowski *et al.*, 2017). Izan ere, `word2vec` eta `GloVe` moduko metodoek hitzen morfologia ez dute kontua izaten, eta hiztegiko berba bakoitzari hitz-bektore bat esleitzen diote. Azken hori muga bat da morfologia aberatseko eleentzat, mota horretako hizkuntzetan maiztasun gutxiko hainbat hitz agertuko baitira, eta, ondorioz, kalitate baxuko hitz-bektoreak izango baitituzte. Gainera, hiztegi handiko hizkuntzetan hitz mailan lan egitea arazoa dela, hiztegitik at dauden hitzen ez baitute errepresentaziorik. Skip-gram eredua oinarri legez hartuta, `fastText` ereduak berba bakoitza karaktereen n-grama multzo legez adierazten du. Esaterako, *bihotz* hitza 3-grama batekin adie-

raziz gero, hurrengo itxura izango luke: `<bi, bih, iho, hot, otz, tz>`. Hala, 3-grama multzo horretako elementu bakoitza hitz-bektore batekin adierazten da, eta *bihotz* hitzaren errepresentazioa sei 3-gramen batura da.

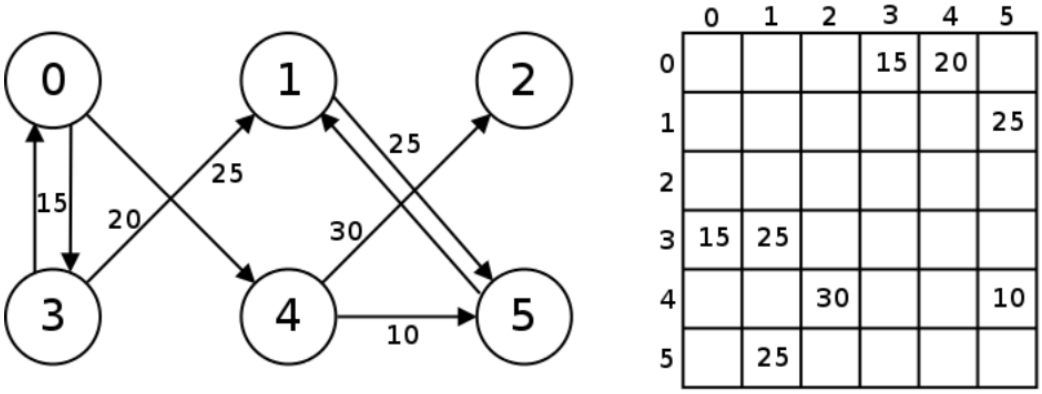
Iragarpen-metodoekin errepresentazio distribuzionalen atala bukatutzat ematen dugu. Hurrengo atalean ikerketa-ildo honetako beste metodo-familia erabilienetako bat deskribatuko dugu, hots, ezagutza-baseetan oinarritutako metodoak.

2.2 Ezagutza-baseetan oinarritutako metodoak

Atal honetan deskribatutako metodoek ezagutza-baseetako informazioa ustiatzen dute. Ezagutza-base terminoa bera oso zabala da eta hainbat sistemei egiten die erreferentzia, baina, horiek guztiek ezagutza esplizituki errepresentatzen dute. Hizkuntzaren prozesamenduan ezagutza-base mota erabilienak sare semantikoak dira, besteak beste, WordNet (Miller, 1995), Freebase (Bollacker *et al.*, 2008), PPDB (Ganitkevitch *et al.*, 2013) edo BabelNet (Navigli and Ponzetto, 2012). Wikipedia entziklopedia askea ere asko erabiltzen da, azken horren artikuluen arteko hiper-estekek ezagutza-base baten egitura osatzen baitute. Kontuan izan, aurreko ataleko iragarpen-metodoek lengoaia naturaleko corpusak prozesatzen dituztela, hots, esplizituki egituratu gabeko baliabideetatik abiatzen direla. Ezagutza-baseak, baina, erlazioen bidez loturiko kontzeptuen sareak dira, eta, ondorioz, azken horien informazio semantikoaren izaera corpusetako edukienaren oso bestelakoa da. Zehazki, corpusetan hitzen arteko erlazioak (inplizituak) ahaidetasunezkoak izateko joera daukate, eta ezagutza-baseetan, gehienbat, antzekotasunezkoak.

Bada, ezagutza-baseetan oinarritutako metodoak aspalditik ezagutzen dira hizkuntzaren prozesamenduaren komunitatean (Mihalcea, 2005; Sinha and Mihalcea, 2007; Agirre and Soroa, 2009; Hassan and Mihalcea, 2011; Witten and Milne, 2008; Banerjee and Pedersen, 2002; Budanitsky and Hirst, 2006) eta azken horiekin lortutako kontzeptuen errepresentazioek berebiziko garrantzia dute hizkuntzaren prozesamenduko hainbat atazatan, hala nola, entitate izendunen desanbiguazioan (Barrena *et al.*, 2016; Lazic *et al.*, 2015; Chisholm and Hachey, 2015) adiera desanbiguazioan (Navigli and Ponzetto, 2012; Agirre *et al.*, 2009b), informazio erauzketan (Banko *et al.*, 2007) eta, noski, antzekotasun semantikoan (Budanitsky and Hirst, 2006; Agirre *et al.*, 2009a).

Metodo horien artean, grafoen teorian oinarritutakoak hizkuntzaren pro-



2.3 irudia – Ezkerraldean, sei erpineko grafo baten adibidea. Grafoa ez-zuzendua da eta erpinak lotzen dituzten ertzek pisuak dituzte. Eskumaldean, 6×6 tamainako grafoaren auzokide-matrizea, erpinek beren auzokideekin (zutabeak) dituzten erlazioez (ertzen pisuez) osatua. Iturria `goo.gl/bWgJhf`.

zesamenduan asko erabiltzen dira, grafoaren ezaugarri estrukturalak bilatzen eta ustiatzen baitituzte. Gainera, grafoaren egitura osoa ustiatzen dute, eta, ondorioz, beren soluzioak globalki optimoak dira. Metodo horiek ezagutza-baseak grafo baten legez ulertzen dituzte, ezagutza-baseko nodoak grafoko erpinak balira bezala tratatuz eta erlazioak ertz bezala. Grafoaren arabera, ertzek norabidea daukate (grafo zuzenduak) edo ez (grafo ez-zuzenduak), eta pisuren bat esleituta izan dezakete. Esaterako, 2.3. irudiak sei erpinez osatutako grafo ez-zuzendu bat azaltzen du, eta bere erpinen arteko erlazioak auzokide-matrize batean daude adierazita.

Grafoen egitura ustiatzeko hainbat metodo daude, batzuk teknika sinpleak erabiltzen dituzte (esaterako, nodoen irteera-mailan oinarritutakoak[11] (Faloutsos *et al.*, 1999)) eta beste batzuk, ordea, konplexuagoak (PageRank (Brin and Page, 1998)) erabiltzen dira. Bada, grafoen teorian oinarritutako algoritmoetatik erabilienetakoa hizkuntzaren prozesamenduan PageRank algoritmoa (Brin and Page, 1998) da, eta ausazko ibilbideetan oinarrituta dago. Jatorrian algoritmo hori bilatzaileetan indexatutako web-orrien sailkapena egiteko pentsatu zen, eta ere eraginkor bezain arrakastatsua izan zen. Algoritmoaren egileek "erabiltzailearen jokabidea" modelatu dute; hau da, erabiltzaileak web-orri batetik hasten da eta beste web-orri batzuetara

[11] Hau da, nodoetatik irteten diren hiper-esteketan oinarritua.

estekaz esteka salto egiten doa, eta, aspertzerakoan, ausazko web-orri bat aukeratu eta beste ibilbide bat hasten du.

Hizkuntzaren prozesamenduan ezagutza-baseen oinarritutako hainbat errepresentazio (Mihalcea, 2005; Sinha and Mihalcea, 2007; Agirre and Soroa, 2009) algoritmo horri (edo bere hedapenen bati) jarraiki kalkulatu dira, hitzen antzekotasun- eta ahaidetasun-atazetan emaitza onak emanez (Agirre *et al.*, 2009a; Camacho-Collados *et al.*, 2016).

Azken urteotan, gainera, ezagutza-baseak hainbat hizkuntzatara hedatu dira (Gonzalez-Agirre *et al.*, 2012). Gainera, BabelNet (Navigli and Ponzetto, 2012), MENTA (de Melo and Weikum, 2010) eta ConceptNet (Speer and Havasi, 2013) modukoak ere osatu dira, eleaniztunak izateaz gain, hainbat ezagutza-baseen informazio estrukturala batzen dutenak. Ezagutza-base horien egitura ustiatuz gero, noski, elearteko atazetarako informazio oso baliagarria daukate. Gauzak horrela, tesi-lan honetan elearteko antzekotasuna atazan erabiliko dugu informazio hori.

Grafoen teorian oinarritutako metodoak alde batera utzita, ezagutza-baseen egitura ustiatzeko aukera gehiago ere badaude. Bada, hurrengo atalean errepresentazio distribuzionalak ezagutza-baseetako informazio estrukturalarekin aberasteko metodoak deskribatuko ditugu, hitz-bektoreen arteko antzekotasun-erlazioak indartzen baitituzte. Ildo horri jarraiki, 5.2.1. atalean guk ere hitz-bektoreen antzekotasuna areagotzeko metodo bat proposatuko dugu.

Hizkuntzaren prozesamenduan ezagutza-base erabilienetakoa WordNet da, eta tesi-lan honetan ere azken hori erabili dugu. Hori dela-eta, hurrengo atalean WordNeten nondik norakoak azalduko ditugu.

2.2.1 WordNet

WordNet ingelesezko ezagutza-basea da (Miller, 1995), ingelesezko hitz eta adieren informazio egituratua duen lexikoia. Azken horretan izenak, aditzak, adjektiboak eta adberbioak *synset* (*synonym set*) izenekoetan antolatzen dira. *Synset* bat sinonimo-multzo bat da, kontzeptu lexikal edo adiera bati dagokiona, hainbat hitzez osatua. Azken horiei *variant* ere deitzen zaie. Esaterako, *car*, *auto* eta *automobile* hitzak *synset* bereko *variantak* dira, adiera bera baitute. Gainera, *synset* bakoitzari glosa bat esleitzen zaio, *synsetaren* definizioa daukana. Esaterako, aipatutako *synsetaren* glosa hurrengoa da: *a motor vehicle with four wheels.*

WordNeteko erlazio garrantzitsuenetakoa sinonimia bada ere, hainbat erlazioz osatua dago. Bada, *synsetak* beren artean sinonimiaz barik, beste erlazio semantiko batzuez lotzen dira, eta azken horien artean garrantzitsuena hiperonimia-hiponimia da. Hiperonimia-hiponimia erlazioak *synset* orokorrenak *synset* zehatzagoekin lotzen ditu. Aurreko adibideari jarraiki, {car, auto, automobile} *synsetaren* hiperonimoa *vehicle* da, eta, aipatutako *synset* horren hiponimoa, ordea, *taxi*.

Egun WordNet hedatzen dabiltza, eta tesi-lan honetan 3.0 bertsioa erabili dugu, 117.659 *synset* eta 525.356 erlazio dituena, 82.115 izen, 13.767 aditz, 18.156 adjektibo eta 3.621 adberbio. Gainera, WordNet beste hainbat hizkuntzatara hedatu da, besteak beste, italierarako, gaztelerarako, frantserako eta euskararako (Vossen, 1998; Pociello *et al.*, 2011). Kontzeptu bera adierazteko, hizkuntza desberdinek *synset* bera erabili dezakete ala ez, eta, azken hori bada kasua, *synseten* arteko mapaketa eskuragarri dago (xehetasun gehiago 5. kapituluan). Eleaniztasun hori medio, Wordnet asko erabiltzen da hizkuntzaren prozesamenduko hainbat atazatan, hala nola, itzulpen automatikoan, informazio-erauzketan eta galdera-erantzun sistemetan. Bada, tesi-lan honetan hitzen arteko elearteko antzekotasunerako erabiliko dugu.

2.3 Testu-corpusak eta ezagutza-baseak uztartzen

Metodo distribuzionalek (ik. 2.1.3. eta 2.1.4. atalak) azaleko informazio linguistikoa darabilte[12], eta, ondorioz, ahaidetasuna eta antzekotasuna nahasteko joera daukate (ik. 1.2. atala). Hala ere, ezagutza-baseetako informazio semantikoa testu-corpusetakoaren desberdina dela kontuan izanik, bi iturri semantiko horien osagarritasuna ustiatzeko eta hitzen errepresentazio hobeak erdiesteko hainbat metodo proposatu dira. Horretarako, metodo horiek testu-corpusetako agerkidetzetako eta ezagutza-baseetako erlazioetako informazioa uztartzen dute.

Hitzen arteko antzekotasun-erlazioak indartze aldera, hainbat autorek errepresentazio distribuzionaletan ezagutza-baseetako informazioa sartu dute (Halawi *et al.*, 2012; Faruqui *et al.*, 2015; Rastogi *et al.*, 2015; Mrkšić *et al.*, 2016, 2017), bi iturburu semantikoetako informazioa uztartuz beren

[12]Esaterako, leiho baten baitako hitz-testuinguru agerkidetzak.

hitz-bektoreetan. Bada, errepresentazio distribuzionaletako hitz-testuinguru agerkidetzetako informazioaz gain, aipatutako eredu horiek WordNet (Miller, 1995), Freebase (Bollacker *et al.*, 2008) edo BabelNet (Navigli and Ponzetto, 2012) bezalako ezagutza-baseetako erlazioen informazioa[13] txertatzen dute hitz-bektoreen ikasketa-prozesuan; hots, testu-corpusetatik at dagoen informazio gehigarria sartzen dute. Hori dela-eta, testu-corpusetan oinarritutako hitz-bektoreei ezagutza-baseen erlazioen murriztapenak sartzen dizkiete eta antzekotasuna areagotzen dute.

Testu-corpusak eta ezagutza-baseak uztartzeko tekniken artean hurrengo bi familiak daude: alde batetik, hitz-bektore aberastuak hasieratik ikasten dituztenak eta ESD baten galera-funtzioan erlazioen informazioa txertatzen dute (Halawi *et al.*, 2012; Wang *et al.*, 2014; Yu and Dredze, 2014; Xu *et al.*, 2014; Bian *et al.*, 2014; Bollegala *et al.*, 2016) bateratze-metodo bezala izendatuko ditugu. Beste aldetik, aurre-entrenatutako hitz-bektoreak ezagutza--baseetako erlazioekin eguneratzen dituztenak fintze-metodo bezala izendatuko ditugu (Rothe and Schütze, 2015; Faruqui *et al.*, 2015; Rastogi *et al.*, 2015; Mrkšić *et al.*, 2016, 2017). Hurrengo bi ataletan bi familia horietako metodoen xehetasun gehiago emango ditugu.

2.3.1 Bateratze-metodoak

Metodo hauen sarrerak testu-corpusak eta ezagutza-baseetako murriztapenak dira, eta, aurretik aipatu bezala, hitz-testuinguru agerkidetzetako eta ezagutza-baseetako erlazioen informazioa uztartzen dute, dena batera. Oro har, iragarpen (Halawi *et al.*, 2012; Wang *et al.*, 2014; Bollegala *et al.*, 2016; Ono *et al.*, 2015) edo kontaketa-metodoren batetik (Schwartz *et al.*, 2015) abiatzen dira, eta azken horren ikasketari ezagutza-baseetako erlazioen informazioa txertatzen zaio. Ezagutza-baseetako informazioa agerkidetzetan oinarritutako ikasketan sartzeko hainbat estrategia erabiltzen dira, hala nola, L2 erregularizazio terminoak (Halawi *et al.*, 2012; Bollegala *et al.*, 2016), hirukoteak (Wang *et al.*, 2014; Xu *et al.*, 2014), konbinazio linealak (Yu and Dredze, 2014), desberdintza sailkapenak (Liu *et al.*, 2015) edo korrelazio kanonikoen analisia bezalako metodo sofistikatuagoak (Osborne *et al.*, 2015).

Kontuan izan, bektore-espazioan txertatutako ezagutza-baseetako erlazioak ezagutza-basearen araberakoak badira ere, metodo horiek gehienbat sinonimiaz (Halawi *et al.*, 2012; Yu and Dredze, 2014) eta antonimiaz (Sch-

[13]Hala nola, sinonimia, antonimia, hiponimia edo hiperonimia.

wartz *et al.*, 2015; Ono *et al.*, 2015) baliatzen direla. Esan gabe doa, antzekotasuna indartzeko erlazioen artean hautagai nagusia sinonimia dela, baina, badirudi antonimiarekin ere efektu beretsuak lortzen direla. Are gehiago, hurrengo ataleko metodoetan biak batera erabiltzen dira, antzekotasun-erlazioak are gehiago sendotuz.

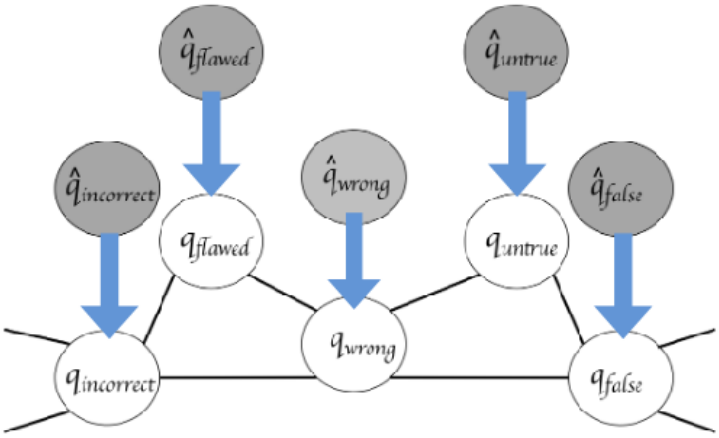
4. kapituluko testu-corpusak eta ezagutza-baseak uztartzeko proposamena CLEARekin konparatu dugu. Gainera, 5.2.1. atalean proposatutako metodoetako bat familia honen baitan kokatzen da. Zehazki, `word2vec` eredu-multzoko Skip-gram ereduari L2 erregularizatzaile bat txertatu diogu (Halawi *et al.*, 2012; Bollegala *et al.*, 2016), baina, gure kasuan, murriztapen elebakarrak barik elebidunak sartzeko.

2.3.2 Fintze-metodoak

Aurreko ataleko ikuspegiak bezala, fintze-metodoek testu-agerkidetzetan oinarritutako errepresentazio distribuzionalak ezagutza-baseetako erlazioez aberasten dituzte. Hala ere, aberastutako hitz-bektoreak hasieratik ikasi beharrean, aurre-entrenatutako testu hitz-bektoreak dira abiapuntua, eta, postprozesu baten bidez, azken horiek ezagutza-baseetako informazio erlazionalarekin fintzen dira.

Familia honetako metodoetan ere aniztasun handia dago, bai erabilitako baliabideen aldetik eta bai estrategien aldetik. Bateratze ikuspegiko joerari jarraiki, abiapuntua iragarpen-metodoekin aurre-entrenatutako testu naturaleko hitz-bektoreak dira (Faruqui *et al.*, 2015; Rothe and Schütze, 2015; Mrkšić *et al.*, 2017; Wieting *et al.*, 2015; Jauhar *et al.*, 2015), baina, baita kontaketetan oinarritutakoak (Faruqui *et al.*, 2015) eta ezagutza-baseekin aberastutakoak (Mrkšić *et al.*, 2016) ere[14]. Ezagutza-baseetako murriztapenak hainbat metodoren bidez txertatzen dira aurre-entrenatutako hitz-bektoreetan, hala nola, kosinu bidezko antzekotasunarekin (Nguyen *et al.*, 2016; Mrkšić *et al.*, 2016, 2017), haztapenen bidezko konbinazioekin (Faruqui *et al.*, 2015; Jauhar *et al.*, 2015) edo metodo korrelazio kanonikoen analisi orokortua bezalako metodo konplexuagoekin (Rastogi *et al.*, 2015). 2.4. irudiak Faruqui *et al.*-en (2015) *retrofitting* metodoaren estrategia azaltzen du; ezagutza-baseko sinonimoen murriztapenen bidez, sinonimo horien testutik erauzitako hitz-bektoreak (grisez) ezagutza-baseko egiturari jarraiki finduko

[14]Salbuespen legez, Recski *et al.*-ek (2016) sei bektore-espazioren erregresioa egiten du, testu naturaleko hitz-bektoreez eta aberastutako hitz-bektoreez baliatuz.



2.4 irudia – *Retrofitting* murriztapenen eragina testu naturaleko bektore-espazioan. Jatorrizko testu naturaleko hitz-bektoreei (grisez) sinonimia murriztapenen bidez ezagutza-baseko egiturako informazioa sartzen zaie eta eguneratu egiten dira (zuriz), beren antzekotasuna handituz. Iturria Faruqui *et al.* (2015).

dira (zuriz) eta, ondorioz, beren arteko antzekotasuna areagotuko da.

Metodo gehienek hitzen errepresentazioekin lan egiten dute, baina, Wieting *et al.*-ek (2015) parafrasien bidez esaldien errepresentazioak kalkulatzea oso eraginkorra dela erakutsi du, eta Rothe and Schütze (2015) lanean hitz-bektoreak modu arrakastatsuan adieretara eta lexemetara hedatu ditu. Aurreko atalean bezala, familia honetan ere sinonimiaren erabilera oso hedatua dago (Faruqui *et al.*, 2015; Rothe and Schütze, 2015), baina, sinonimiaren eta antonimiaren konbinaketarekin antzekotasun-erlazioak are gehiago indartu dira (Mrkšić *et al.*, 2016, 2017).

Tesi-lan honen 4. kapituluko testu-corpusak eta ezagutza-baseak uztartzeko proposamena *retrofittingekin* eta Rastogi *et al.*-en (2015) lanarekin konparatu dugu.

2.4 Elearteko espazioak

Hitz-bektore elebakarrek hizkuntzaren prozesamenduan izandako arrakastak dela-eta, azken horiek elearteko atazetan erabiltzeko interesa ere piztu da. Gauzak horrela, hainbat hizkuntzetako hitzen esanahiak elearteko bektore-

espazio bateratuetan kodetzeko metodo eta estrategia anitz agertu dira. Izan ere, Ruder-ek (2017) dioen bezala, elearteko hitz-bektoreak oso erakargarriak dira: lehenik, hitzen esanahien inferentzia testuinguru eleaniztunen bitartez burutzen denez, elearteko antzekotasuna ahalbidetzen dutelako, hainbat atazetarako hain garrantzitsua dena[15]; bigarrenik, elearteko hitz-bektoreek hizkuntzen arteko informazio transferentzia ahalbidetzen dutelako. Elearteko transferentzia horrek are garrantzia gehiago hartzen du baliabide handiko ele baten eta gutxitu baten artean denean, azkenengoari onura egingo baitio.

Mapaketetan oinarritutako metodoak

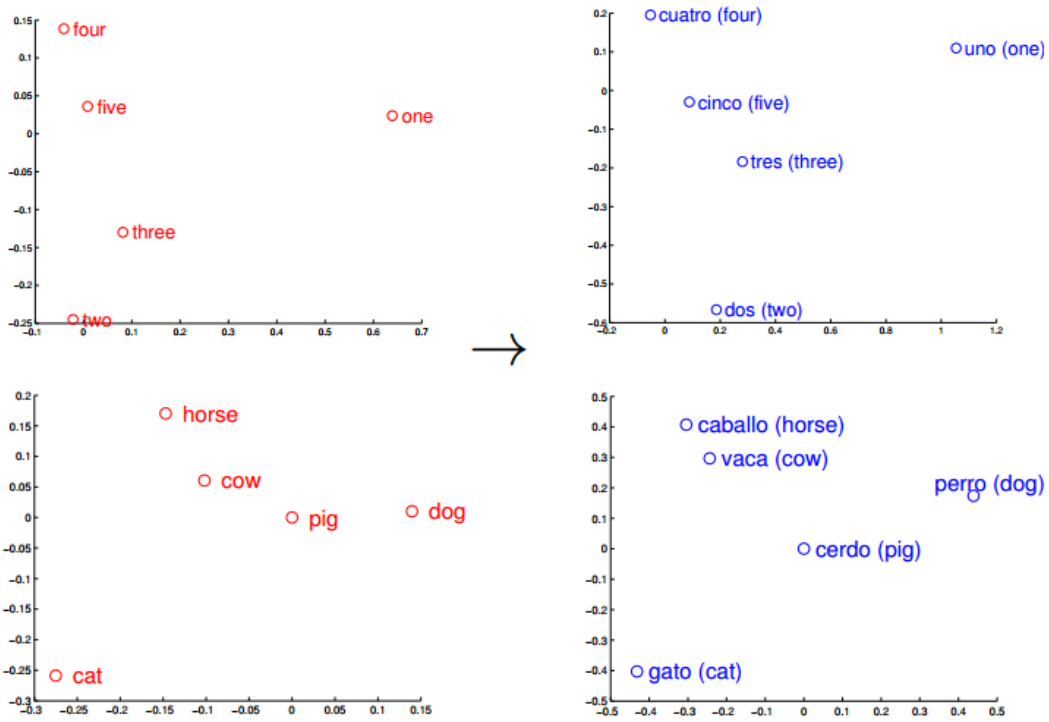
Metodo hauek testu-corpusetatik erauzitako bektore-espazio elebakar bereizietatik abiatzen dira, eta, hiztegi elebidun baten oinarrituta, espazio batetik besterako transformazioa ikasten dute. Mapaketak egiteko espazio elebakarren artean isomorfismoa dagoela hartzen da hipotesitzat; hau da, bereizita ikasi badira ere, espazio elebakarrek egitura berdina izango dutela. Teknika honen aitzindaria Mikolov *et al.* (2013b) izan zen, distantzia euklidearren karratuak minimizatzeko transformazio lineal bat aplikatu ziena espazio elebakarrei. Azken horren optimizazio ereduari hainbat hobekuntza aplikatu dizkiote; Zhang *et al.*-ek (2016) transformazio-matrizea ortogonala izatera derrigortzen du, Xing *et al.*-ek (2015) ortogonalitateaz gain hitz-bektoreen normalizazioa eta (distantzia euklidearraren ordez) kosinua sartzen ditu, eta Lazaridou *et al.*-ek (2015) *zero-shot* ikasketa[16] erabili du. Beste estrategia bati jarraiki, Faruqui and Dyer-ek (2014) korrelazio kanonikoen analisiaren bidez bi bektore-espazio elebakar espazio partekatu batean mapatzen dute, eta Lu *et al.*-ek (2015) azken horren bertsio ez lineala proposatzen dute. Artetxe *et al.*-ek (2016) azken horiek guztiak orokortu zituen, soluzio zehatza lortzeko metodo eraginkorra proposatuz[17].

2.5. irudia paragrafo honetan aipatutako isomorfismoaren adibide bat da. Bada, irudi horrek independeteki ikasitako ingelesezko eta gaztelerazko bektore-espazioak erakusten ditu; zehazki, zenbakien eta animalien hitzbektoreak, bi dimentsiotako espazio batean. Adibide horretan ikusten denez, bereizita ikasi badira ere, hizkuntza desberdinetako bektore-espazioek antz-

[15]Lexiko-inferentzia elebiduna eta elearteko informazio erauzketa, besteak beste.

[16]Klase anitzetako sailkapena egiten da, baina, hainbat klaseren inguruko entrenemendurako daturik gabe.

[17]Aurreko metodoek gradiente-jaitsisiera estokastikoa erabiltzen dute parametroen optimizazioan.



2.5 irudia – Ingelerazko (ezkerrean) eta gaztelerazko (eskuman) zenbakien eta animalien errepresentazio distribuzionalak bi dimentsiotan irudikatuta. Irudiak bi hizkuntzetako bektore-espazioek antz-antzeko egitura dutela erakusten du, hitzen arteko distantziak eta posizioak mantendu egiten baitira hizkuntzen artean. Iturria Mikolov *et al.* (2013b).

antzeko egitura daukate, hots, hitz-bektoreek distantzia eta posizio bera daukate beren artean, isomorfoak dira. Isomorfoak izanik ere, bektore-espazio bereiziak mapatu egin behar dira eleen arteko hitz-bektore baliokideak bata bestearen "gainean" egon daitezen.

Aurrekoa metodoari lotuta, hiztegi elebidunen erabilera murrizteko edo erabat ezabatzeko hainbat estrategia proposatu dira. Esaterako, Zhang *et al.*-ek (2016) hiztegiaren sarrerak hamarrera arte murrizten ditu, baina, lexiko-indukzio elebiduna moduko atazan zorrotzetan ez du emaitza onik. Artetxe *et al.*-en (2017) lanean, abiapuntu bezala hogeita bost sarreratako hiztegi bat edo zenbakiak hiztegi bezala erabilita, iterazioetan oinarritutako metodo bidez mapaketarako hiztegia inferitzea lortu dute, mapaketa modu arrakastatsuan burutuz.

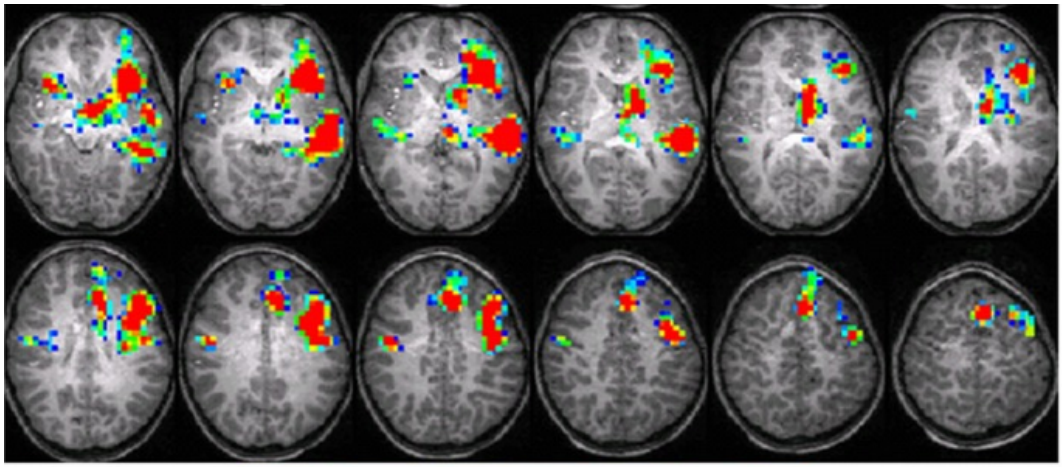
Artetxe *et al.* (2018) eta Conneau *et al.* (2017) haratago doaz, eta hiztegi elebiduna inongo lagin elebidun barik inferitzen dute. Artetxe *et al.*-ek (2018) are gehiago hobetu du Artetxe *et al.* (2017) laneko iterazioen metodoa, hiztegiaren erabilera guztiz ezabatuz, eta, ondorioz, modu guztiz ezgainbegiratuan. Conneau *et al.*-en (2017) mapaketa-metodoa bi corpus elebakarretatik abiatzen da, eta, entrenamendu aurkariduna erabiliz, bektore-espazio batetik besterako mapaketa ikasten doa eta eleen arteko hiztegia inferitzen.

Artetxe *et al.* (2016) mapaketa-metodoa garrantzitsua da tesi honetan, 5. kapituluan gure metodo elebidunaren oinarri-lerroa izango delako.

2.5 Ebaluazioa

Tesi-lan honetan jarraitutako ikerketa-ildoan hitzen errepresentazioen ebaluazioak garrantzi handia dauka (Griffiths *et al.*, 2007; Baroni *et al.*, 2014), eta ebaluazio estrinsekoetan eta ebaluazioa intrinsekoetan (Schnabel *et al.*, 2015) banatzen da. Ebaluazio estrinsekoek hitz-bektoreak aplikazio baten sarrera legez erabiltzen dituzte, eta azken hori hizkuntzaren prozesamenduko ataza errealen batean probatzen da. Zenbat eta portaera hobea azaldu hitzen errepresentazioek ataza horretan, orduan eta kalitate hobekoak izango dira. Ataza horien artean entitate izendunen ezagutzea (Turian *et al.*, 2010; Collobert *et al.*, 2011), sentimendu analisia (Schnabel *et al.*, 2015), rol semantikoen etiketatzea (Collobert *et al.*, 2011) edo testu-loturen detekzioa (Bowman *et al.*, 2015) daude.

Ebaluazio intrinsekoetan, ordea, hitzen errepresentazioen emaitzak giza irizpideekin konparatzen dira, zehazki, gizakiek emandako hitzen erlazioen irizpideekin. Eskuz sortutako hitz-multzoei giza irizpideak esleitzen zaizkie, eta azken horiek batzeko zuzenean parte-hartzaile batzuek egin dezakete, ala *Mechanical Turk* bezalako plataformen bidez lortu daitezke. Ebaluazio intrinsekoak bi taldetan banatzen dira: kontzienteak eta subkontzienteak. Ebaluazio intrinseko kontzienteak dira bietatik erabilienak, eta giza irizpideak pertsonen prozesu kontzienteen bidez jasotzen dira; hau da, parte-hartzaileek erantzunak pentsatu behar dituzte. Familia horren barruan hitzen arteko antzekotasuna (Agirre *et al.*, 2009a; Baroni *et al.*, 2014), analogia (Turian *et al.*, 2010; Baroni *et al.*, 2014; Mikolov *et al.*, 2013d), sinonimo-detekzioa (Baroni *et al.*, 2014), kontzeptu-kategorizazioa (Baroni *et al.*, 2014) eta sinonimo-detekzioa (Baroni *et al.*, 2014) daude. Bigarren taldeak gure subkontziente-



2.6 irudia – fMRI teknikaren bidez haurren aditzen sorkuntzan jasotako neurona-aktibazioak. Iturria `https://goo.gl/XZMN9V`.

ko erantzunekin osatutako ebaluazio metodoak daude, eta azkenaldion arreta gehiago jartzen hasi zaie. Bada, talde horretan neurona-aktibazioetan oinarritutakoak (Xu *et al.*, 2016), begien mugimenduetan datuekin osatutako urre-patroiak (Luke and Christianson, 2017; Cop *et al.*, 2017) sartzen dira. Neurona-aktibazioetan oinarritutakoen artean *functional magnetic resonance imaging* (fMRI) eta elektorenzefalogramen (EEG) tekniken erabilera oso hedatuta dago. 2.6. irudiak, esaterako, haurren aditzen sorkuntzan fMRI bidez jasotako neurona-aktibazioak agertzen dira. Neurona-aktibazio horiek urre-patroiak osatzeko informazio baliagarria daukate.

Tesi-lan honetan ebaluazio intrinseko kontziente batez baliatu gara, hitzen arteko antzekotasunaz, hizkuntzaren prozesamenduan tradizio handia daukana. Hurrengo ataletan ebaluazio ataza horren nondik norakoak deskribatuko ditugu.

2.5.1 Hitzen arteko antzekotasun eta ahaidetasun urre-patroiak

Tesi-lan honen sarrerako 1.1. atalean antzekotasun- eta ahaidetasun-erlazioak bereizi ditugu; aurrenekoak antzeko ezaugarriak dituzten hitzen arteko erlazioak dira (*otso* eta *txakur*), eta, bigarrenak, asoziaziozko erlazioren bat duten hitzen artekoak (*otso* eta *ilargi*). Erlazio horiek antzekotasun eta

ahaidetasun urre-patroietan modu kuantitatibo batean islatzen dira, eta gure eredu konputazionalen errepresentazioen kalitatea neurtzeko baliabide oso garrantzitsuak dira. Urre-patroiak gizakiek emandako balioez osatuta daude, eta, ondorioz, gure eredu konputazionalen hitzen errepresentazioek giza irizpideetan oinarritutako antzekotasuna edota ahaidetasuna erreproduzitzeko gaitasuna neurtzeko balio dute.

Azken hori burutze aldera, lehenik, hitzen errepresentazioak kalkulatuko ditugu, eta, errepresentazio horietatik abiatuta, hitzen arteko antzekotasunak. Esan bezala, eredu konputazionalen errepresentazioak ebaluatzeko ereduaren antzekotasun-balio horiek giza irizpideekin duten korrelazioa kalkulatuko dugu. Hala, eredu konputazionalak zenbat eta korrelazio altuagoa izan gizakien erantzunekin, orduan eta kalitate hobea izango dute errepresentazioek. Lan honetan erabilitako korrelazioaren eta antzekotasunaren nondik norakoak hurrengo atalean deskribatzen ditugu.

Gure lanari dagokionez, 3. eta 4 kapituluetan ingelesezko hainbat ahaidetasun eta antzekotasun datu-multzo elebakarrekin ebaluatuko ditugu gure errepresentazioak. 5. kapituluan errepresentazio elebidunekin lan egingo dugunez, azken horien ebaluaziorako hainbat elearteko datu-multzo osatuko ditugu.

Urre-patroi elebakarrak

Urre-patroiak hitz bikoteez osatuta daude, eta bikote bakoitzak antzekotasun- edo ahaidetasun-balio bat dauka esleituta, hainbat partaideren erantzunekin osatua. Urre-patroi horiek osatze aldera, aurretiaz azken horien izaera semantikoaren araberako antzekotasun edo ahaidetasun irizpideak definitu behar dira. Kontua izan, urre-patroietan hitzen antzekotasun- eta ahaidetasun-erlazioak islatzeko irizpideak desberdinak direla. Antzekotasuneko urre-patroien artean erabilienetakoa SimLex999 (Hill *et al.*, 2015) da, baina, RG (Rubenstein and Goodenough, 1965) ere aspaldidanik erabiltzen da. Ahaidetasunekoen artean WordSim353 (Finkelstein *et al.*, 2001) da erabiliena, baina, horrez gain, aipatzekoak dira MTURK287 (Radinsky *et al.*, 2011) eta MEN (Bruni *et al.*, 2014) ere.

Esaterako, hurrengo adibidean SimLex999 zinezko antzekotasun urre-patroiko lau hitz bikote azaltzen dira:

```
clothes closet 3.27
sunset sunrise 2.47
child adult 2.98
cow cattle 9.52
```

Adibidean ikusten den legez, partaideek *cow-cattle* pareari soilik esleitu diote antzekotasun-balio altua, hiponimo/hipernonimo erlazioa baitaukate. Beste hiru pareek, ordea, ahaidetasun-erlazioak dituzte[18], eta, ondorioz, antzekotasun-balio baxuak. Aurreko adibidean ikusten dugunez, *clothes-closet* pareak SimLex999 antzekotasun datu-multzoan 3.27 balioa dauka, baina WordSim353 datu-multzoan antzekotasuna eta ahaidetasuna ez dira bereizten irizpideetan, eta pare horrek 8.00 balioa du esleituta. Bi datu-multzoetan eskalak 0tik 10era doazela kontuan izanik, pare horrek balio nahiko baxua dauka SimLex999 datu-multzoan, eta altua WordSim353n. Izan ere, aurrenekoa osatzeko zinezko antzekotasun irizpideak esplizituki definitu dira, eta azkenengoan, ordea, antzekotasun eta ahaidetasun irizpideak nahastuta daude[19].

Irizpideez gain, urre-patroiaren arabera, bikoteetako hitzak kategoria-gramatikal (izen, adjektibo, aditz) eta abstrakzio maila desberdinetakoak izan daitezke, eta antzekotasun- edota ahaidetasun-balioen eskalak ere aldatzen dira. Eduki aldetik, oro har, oso orokorrak dira, baina azken urteotan domeinu espezifikoetakoak (hala nola, medikuntza (Chiu *et al.*, 2018)) ere azaltzen dabiltza. Esan gabe doa gehiengo nagusia ingelesez dela, baina jatorrizko datu-multzoak beste hizkuntza batzuetara (Camacho-Collados *et al.*, 2015; Mrkšić *et al.*, 2017) ere egokitu dira. Besteak beste, lan honekin, zentzu horretan ere bere ekarpena egin dugu euskal hiztunen irizpideekin osatutako bi urre-patroi (WordSim353 eta RG) euskaratu baitugu (ik. 5. atala).

Elearteko urre-patroiak

Hizkuntzaren prozesamenduan hitzen antzekotasun elebakarra atazarik ezagunenetakoa bada ere (Agirre *et al.*, 2009a; Baroni *et al.*, 2014), azken urteotan errepresentazio eleaniztunek izandako gorakada dela-eta, elearteko urre-patroiak osatzeko proposamenak ere azaldu dira (Kennedy and Hirst, 2012;

[18] *Clothes-closet* bikoteak asoziazio funtzionala dauka, eta *sunset-sunrise* eta *child-adult* pareak antonimo bezala uler daitezke.

[19] Ikusi `https://www.cl.cam.ac.uk/~fh295/simlex.html` eta `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/`, hurrenez hurren.

Camacho-Collados *et al.*, 2015). Gauzak horrela, Camacho-Collados *et al.*-ek (2015) elearteko hainbat datu-multzo eta azken horiek modu automatikoan eta estandarizatuan osatzeko tresna eskuragarri jarri dituzte[20]. Tresna hori lerrokatutako datu-multzo elebakarretatik abiatzen da, eta azken horiek konbinatuz elearteko datu-multzoa sortzen du; zehazki, lerrokatutako hitz bikote baliokide bakoitzerako, elearteko bikoteak sortzen ditu eta antzekotasun- eta ahaidetasun-balio elebakarren batezbestekoak kalkulatzen ditu (xehetasun gehiago Camacho-Collados *et al.* (2015) lanean). Hurrengo adibidean, SimLex999 datu-multzoaren ingeles-italiera bertsioaren lau hitz bikote agertzen dira:

```
clothes armadio 4.21
zone area 8.55
become fare 3.655
pupil studente 9.25
```

Kontuan izan hizkuntzen artean adiera eta konnotazio kulturalen eragina dela-eta, hitz pare berak hizkuntza desberdinetan antzekotasun-balio desberdina izango duela. Esaterako, Ingelesezko SimLex999 urre-patroian *clothes-closet* pareak 3.27 balioa dauka, eta italierazko *vestiti-armadio* baliokideak, ordea, 5.15[21]. Hortaz, elearteko urre-patroi horiek antzekotasunaren bi aspektu jasotzen dituzte: alde batetik, urre-patroiko erlazio semantikoak, (antzekotasun-, ahaidetasun-erlazioak edo biak nahastua), hizkuntzaren independenteak; beste aldetik, elearteko konnotazioek antzekotasun- eta ahaidetasun-irizpideetan duten eragina. Esan gabe doa, hitzen errepresentazioak urre-patroi horiekin ebaluatuz gero, aipatutako bi aspektu horiek jasotzeko gaitasuna ebaluatzen dugula.

Korrelazioa

Atal honen hasieran ebaluazio intrinsekoetan hitzen errepresentazioen emaitzak giza irizpideekin konparatzen direla aipatu dugu. Azken hori gauzatze aldera, hizkuntzaren prozesamenduan datu-multzoen arteko korrelazio-koefizientea erabiltzen da , gure ikerketa-ildoan Spearman (Spearman, 1904) edota Pearson (Galton, 1889) korrelazioak dira ohikoenak. Kontuan izan,

[20] `http://lcl.uniroma1.it/similarity-datasets/`

[21] Balio hori modu automatikoan kalkulatu da, baina baliagarria zaigu erreferentzia legez.

korrelazioa kalkulatu aurretik, lehenik eta behin, urre-patroietako hitz bikoteen antzekotasunak kalkulatzen direla. Oro har, hitzen errepresentazioen arteko antzekotasuna kosinuaren formularen bitartez kalkulatzen da. Datu-multzoko bikote guztientzat hitzen errepresentazioen arteko antzekotasuna kalkulatu ondoren, azken horren balioen eta (giza irizpideekin osatutako) urre-patroiaren balioen arteko korrelazioa kalkulatzen da.

Bada, Pearsonen r korrelazio-koefizienteak X eta Y bi aldagai jarrairen arteko erlazioaren linealtasuna ebaluatzen du. Pearsonen r koefizienteak $[-1, +1]$ tarteko balioak hartzen ditu; -1 balioak X eta Z guztiz linealak direla esan nahi du, baina, linealtasuna negatiboa dela; 1 balioak guztiz linealak direla eta linealtasuna positiboak dela adierazten du; 0 balioak ez dagoela inongo linealtasun erlaziorik; erdibideko beste balioek linealtasun maila adierazten dute. (2.1) ekuazioak Pearsonen r koefizientea deskribatzen da:

$$r = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}} \tag{2.1}$$

Ekuazio horretan, n aldagai jarraiko lagin kopurua da[22], X_i eta Y_i laginen balioak , eta $\overline{X}$ eta $\overline{Y}$ bi datu-multzoetako batez bestekoen balioak. Bada, (2.1) ekuazioko goiko terminoa X eta Y aldagaien arteko kobariantza da eta azpikoak aldagaien desbideratze estandarrak dira.

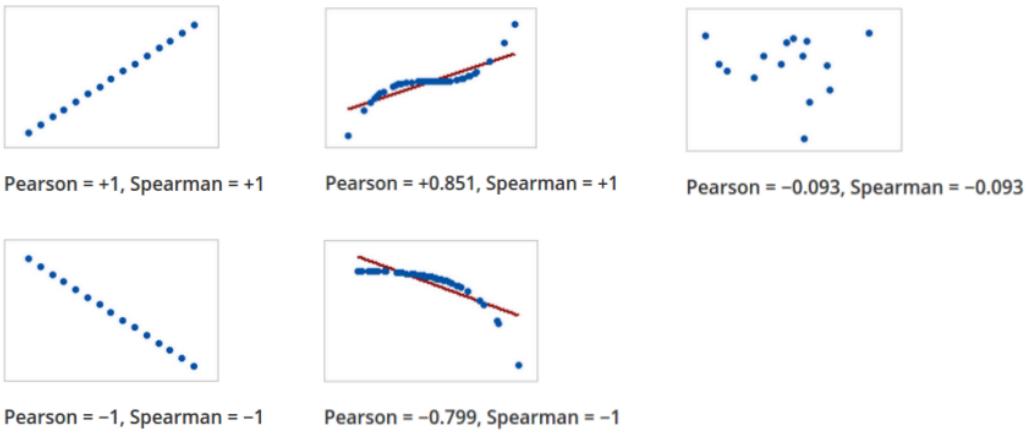
Spearman ρ korrelazio-koefizienteak, ordea, bi aldagai jarrairen arteko erlazio monotonikoak ebaluatzen ditu, zehazki, datu-multzoetako antzekotasunen sailkapen-balioak[23] ebaluatzen ditu. (2.2) ekuazioak ρ koefizientea deskribatzen du:

$$\rho \;=\; 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}, \qquad di = (s(X_i) - s(X_i)) \tag{2.2}$$

Ekuazio horretan $s(X_i)$ eta $s(Y_i)$ X_i eta Y_i laginen sailkapen-balioak dira, hurrenez hurren. Hala, ρ koefizientea bi aldagairen sailkapen-balioen Pearson koefizientea da, eta, azken horren moduan, [-1,+1] arteko balioak hartzen ditu. ρ koefizientearen interpretazioa Spearmanen r koefizientearen antzekoa da, hau da, zenbat eta altuagoa izan, orduan eta antzekotasun gehiago datu-multzoen sailkapenen artean, eta, alderantziz. Hemen ere, $\rho = 0$ emaitzak,

[22]Hau da, datu-multzoetako hitz-bikote kopurua.

[23]Hau da, antzekotasunaren araberako hitz-pareen sailkapen-balioa datu-multzoan (*ranking*).



2.7 irudia – Pearson eta Spearman korrelazio-koefizienteen adibideak. Iturria `goo.gl/FS434M`.

noski, bi datu-multzoen sailkapen-balioen artean antzekotasunik ez dagoela esan gura du.

2.7. irudian Pearsonen r eta Spearmanen ρ korrelazio-koefizienteen balioen hainbat adibide agertzen dira. Irudietako bi ardatzak X eta Y aldagaien balioak dira. Ezkerreko bi irudiek bi koefizienteen muturreko balioak erakusten dituzte; hau da, X eta Y aldagaiek korrelazio maximoa dute r eta ρ koefizienteetan, baina goiko irudian korrelazioa positiboa da eta behekoan negatiboa. Erdiko bi irudiek erdibideko emaitzak erakusten dituzte, bi aldagaien erlazioa ez baita guztiz lienala ($r < 1$), baina, beren arteko sailkapen-balioen berdina da ($\rho = 1$). Eskumako irudiak korrelaziorik gabeko aldagai bi deskribatzen ditu.

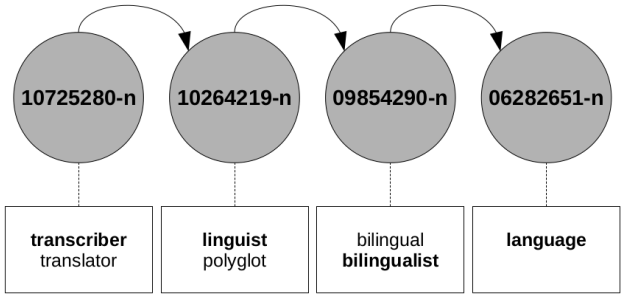
3

Ezagutza-baseetan oinarritutako teknikak: ausazko ibilbideak

Kapitulu honetan, tesi-laneko aurreneko ekarpena deskribatuko dugu, ezagutza-baseetan oinarritzen dena. Ausazko ibilbideez baliatuta, metodo horrek ezagutza-baseetatik corpus sintetikoak erauztea ahalbidetzen du, eta, ondoren, ezagutza-baseetako informazio estrukturala bektore-espazio batean kodetzen. Esperimentuen atalean, ezagutza-basetik erauzitako errepresentazio trinkoak inplizituki ezagutza-baseetako informazio estrukturala gordetzen dutela enpirikoki egiaztatuko dugu.

Beste berba batzuetan esanda, ezagutza-baseetako egitura semantikoa dimentsionaltasun gutxiko bektore-espazio jarrai batean kodetuko dugu. Ezagutza-baseetan oinarritutako teknika ohikoekin alderatuta (ik. 2.2. atala), gure metodoak informazio bera bektore trinkoagoetan kodetzeko gai izango da[1]. Abantaila horretaz gain, hitzen antzekotasun-atazan emaitza beretsuak edo hobeak lortuko ditugu, eta modu azkarragoan. Gainera, ezagutza-baseetan oinarritutako gure metodoa testuetan oinarritutakoekin konbinatu eta hitzen antzekotasun-emaitzak are gehiago hobetu ditugu, bi metodoek informazio semantiko osagarria dutela ondorioztatuz.

[1]Guk proposatutako WordNeteko hitzen errepresentazio trinkoek 300 dimentsio dituzte, eta ohiko metodoetako bektoreek, ordea, ehunka milako.



3.1 irudia – Ingelesezko Wordnet-aren gaineko ausazko ibilbide elebakarra. Ibilbideak zeharkatutako kontzeptuak (*synsetak*) grisez, eta azken horien lexikalizazioak laukien barruan. Ausaz aukeratutako lexikalizazioak urrats bakoitzean letra lodiz.

3.1 Metodoa

Atal honetan lan honen ekarpen nagusienetakoak izango ditugu hizpide; hots, ezagutza-baseetako corpus sintetikoak eta azken horien errepresentazio trinkoak, hurrengo bi kapituluetako metodoen euskarriak. Izan ere, ezagutza-baseetako corpus sintetikoek eta azken horien errepresentazio trinkoek testu-corpuseko informazioarekin hainbat konbinazio egitea ahalbidetuko digute (ik. 4. kapitulua), eta, gainera, (WordNeteko corpus sintetikoak zein errepresentazio trinkoak) ele anitzetara hedatzeko errazak dira (ik. 5 kapitulua).

Corpus sintetikoak erauzteko edozein ezagutza-basez baliatu badaiteke ere, kapitulu honetan 2.2.1. atalean aipatutako glosadun WordNet 3.0 (Miller, 1995) erabili digu.

WordNet corpus sintetikoak eta errepresentazio trinkoak lortzeko bi pausu jarraitu behar dira; lehenik, ausazko ibilbideetan oinarritutako algoritmo baten bidez WordNet corpus sintetikoa osatu; gero, corpus horren errepresentazio trinkoak kalkulatu. Hala, hurrengo bi ataletan corpus sintetikoak osatzeko algoritmoa eta, azken horietatik abiatuta, errepresentazio trinkoak kalkulatzeko iragarpen-metodoaren nondik norakoak deskribatuko ditugu.

3.1.1 Ausazko ibilbide elebakarrak

Ezagutza-baseetako testuinguru sintetikoak sortze aldera, gure metodoa ezagutza-baseen gainean egindako ausazko ibilbideez baliatzen da. Esaterako,

3.1. irudian WordNet gainean burututako hurrengo ausazko ibilbidea agertzen da: *transcriber linguist bilingualist language*. Ibilbideko urrats bakoitzean WordNeteko egituran erlazionatutako kontzeptu batera egiten du jauzi, eta kontzeptu horri esleitutako lexikalizazioetako bat ausaz aukeratzen du. Halako ibilbide bakoitza corpus sintetikoko lerro bat da.

Aipatutako ausazko ibilbideen metodoarekin sortutako testuinguruak fitxategi batean gorde eta corpus sintetiko bat osatzen dute. Ondoren, neuronasare batek corpus arrunta balitz bezala prozesatzen du, eta errepresentazio trinkoak sortzen ditu.

Bada, $G = (K, E)$ ezagutza-basea grafo ez-zuzendu baten moduan ulertuko dugu, non K erpinak (kontzeptuak) eta E kontzeptuen arteko ertzak (ezagutza-baseko erlazioak) diren. Bada, $N(k)$ k kontzeptuaren inguruko auzokideak dira grafoan (ertz baten bidez zuzenean lotutako erpinak), eta $L(k) = w_1, ..., w_n$ k kontzeptuaren lexikalizazioak. Kapitulu honetako eta 4. eta 5. kapituluetako esperimentuetan, WordNeteko *synsetak* erpin legez erabiltzen ditugu[2], eta *synseten* arteko erlazioak ertz moduan. WordNeteko erlazio oro hartzen ditu kontuan, glosa erlazioak[3] barne. Azken horiek glosatutako *synsetaren* eta glosako *synset* ororen arteko erlazioak dira. Lexikalizazioen multzoa *synseten* lemez osatuta dago.

1. taulak ausazko ibilbideen algoritmoa azaltzen du. Ausazko ibilbideak PageRank burutzeko Monte Carlo metodo batez (Avrachenkov *et al.*, 2007) baliatzen da. Bada, algoritmoa grafoko ausazko erpin batetik abiatzen da. Ibilbidearen urrats bakoitzean probabilitate bat jaurtitzen du algoritmoak: α probabilitatearekin nodo auzokide bat aukeratzen da, ertza ausaz aukeratuta (aukera orok distribuzio uniformea dauka); $1 - \alpha$ probabilitatearekin ibilbidea geratu egiten da. Kontzeptu bat aukeratzen den guztietan, ausaz kontzeptuko lexikalizazio bat emititzen. Ibilbidea geratzen den orotan, lexikalizazio multzo bat sortzen da, eta ibilbidean emititutako hitzak fitxategi batean inprimatzen dira. Aurretiaz definitutako I testuingururaino heltzerakoan, algoritmoak ibilbideak burutzeari uzten dio. Beraz, algoritmoak bi hiper-parametro ditu, α eta I.

3.2 taulak glosadun WordNet 3.0 grafotik erauzitako hurrengo testuinguru sintetikoaren xehetasunak azaltzen du: *paw feline feline felid ounce panthera lion sekhet*. Algoritmoak ausaz *02439929-n synsetetik* abiatzen da, eta *paw*

[2]WordNeten kontzeptuei *synset* legez egiten zaie erreferentzia. Lan honetan modu berean erabiltzen ditugu termino biak.

[3]`http://wordnet.princeton.edu/glosstag.shtml`

Algorithm 1 Ausazko ibilbide elebakarrak

Sarrera: K kontzeptu multzoa
$N(k)$, $k \in K$ kontzeptuaren auzokideak grafoan
$D(k)$, $k \in K$ kontzeptuaren lexikalizazioak
I, testuinguru sintetikoen kopurua
α, moteltze-faktorea
Irteera: CS, corpus sintetiko elebakarra
$CS \leftarrow []$
$i \leftarrow 0$
repeat
  $S \leftarrow []$
  $k \in K$ erpina aukeratu, $1/|K|$ probabilitatearekin
  repeat ▷ Ibilbidea jarraitu
    $w \in D(k)$ hitza aukeratu, $1/|D(k)|$ probabilitatearekin, S-n sartu
    $k' \in N(k)$ erpina aukeratu, $1/|N(k)|$ probabilitatearekin
    $k \leftarrow k'$
  until $random() > \alpha$ ▷ Ibilbidea geratu
  $CS = CS \cup S$ ▷ Testuinguru berria
  $i \leftarrow i + 1$
until $i == I$

3.1 taula – Monte Carlo metodoetan oinarritutako ausazko ibilbide elebakarren algoritmoa. Grafoan erlazionatutako erpinak zeharkatzen ditu, eta urrats bakoitzean lexikalizazio bat emititzen. Ibilbide bat bukatzen, fitxategi batean gorde eta beste ibilbide bat hasten du.

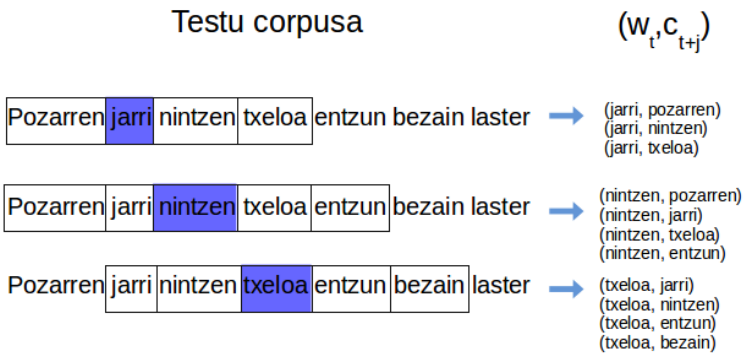
synsetak	lexikalizazioak	erlazioa
02439929-n	**paw**	meronym part^{-1}
02120997-n	felid **feline**	related to
02881888-a	**feline**	related to
02120997-n	**felid** feline	gloss^{-1}
02128757-n	panthera_uncia **ounce** snow_leopard	meronym member^{-1}
02128120-n	**panthera** genus_panthera	gloss
02129165-n	**lion** panthera_leo king_of_beasts	gloss^{-1}
09513430-n	**sekhet**	—

3.2 taula – Ausazko ibilbide elebakarren adibidea. Ezkerreko zutabean *synsetak*, erdikoan *synseten* lexikalizazio posible oro eta eskuman ibilbideko hurrengo urratserako WordNet erpina. $^{-1}$ ikurrak alderantzizko erlazioa adierazten du, grafoa ez-zuzendua baita. Synset bakoitzean ausaz aukeratutako lexikalizazioak letra lodiz.

emititzen du. Ondoren, *02120997-n synsetera* jauzi egiten du *meronym*$^{-1}$ erlazioaren bitartez (meronimoaren alderantzizkoa), eta bi lexikalizazio posibleetatik *feline* aukeratzen du. Hurrengo urratsean *02881888-a synsetera* doa *relate to* erlazioaren bidez, eta aurreko pausuko *feline* lexikalizazioa dagoenez aukera bezala, azken hori birritan jarraian emititzen du. Gero, algoritmoak *felid*, *ounce*, *panthera*, *lion* eta *sekhet* emititzen ditu, eta, azkenik, geratu egiten da.

Hala, aipatutako adibidearen moduko ibilbide bat bukatzerakoan, algoritmoak fitxategi batean inprimatzen du. Horren ondoren, grafoko beste nodo bat ausaz aukeratzen du eta beste ibilbide batekin hasten da. I ibilbide sortu dituenean, geratu egiten da.

Behin testuinguru sintetikoz osatutako corpus sintetikoa esku artean izanik, 2.1.4. aipatutako Skip-gram ereduko sarrera legez erabiliko dugu, eta testu-corpus arrunta balitz bezala prozesatuko. 2.1.4. atalean azaldu bezala, testuetan oinarritutako metodoen bitartez, grafoko lexikalizazioen errepresentazio trinkoak lortuko ditugu; hots, grafoko informazio estrukturala bektore-espazio trinko batean kodetuko dugu. Corpus sintetikoak sortzeko algoritmoa azaldu ondoren, hurrengo atalean Skip-gram ereduaren xehetasunak izango ditugu hizpide.



3.2 irudia – Skip-gram ereduak testu-corpusak prozesatzeko erabilitako estrategia. Leiho irristakor batekin (laukiez osatutakoa) corpus osoa zeharkatzen du, eta behatutako hitzak (lauki urdina) bere testuingurukoak (lauki zuria) aurresateko erabiltzen ditu. Eskubian hitz-testuinguru bikoteak azaltzen dira, hitzak w_t zutabean eta testuinguruak c_{t+j} zutabean.

3.1.2 WordNeteko hitzen errepresentazio trinkoak kalkulatzeko metodoa

Aurreko atalean esan bezala, behin WordNeteko corpus sintetikoa osatuta, azken hori 2.1.4. atalean azaldutako iragarpen-metodo baten bidez prozesatu eta WordNeteko lexikalizazioen errepresentazio trinkoak kalkulatuko ditugu, hau da, WordNeteko informazio estrukturala bektore-espazio batean kodetuko dugu. Tesi-lan honetan zehar `word2vec`[4] eredu-multzoko Skip-gram eredua erabili dugu WordNeteko errepresentazio trinkoak kalkulatzeko. Aurrerago ikusiko dugunez, kapitulu honetan CBOW ereduaz ere baliatu gara errepresentazio trinkoak lortzeko, baina lan honetan ez dugu gehiagorik erabili.

3.2. irudiak Skip-gram ereduaren metodoa erakusten du; corpusa hitzez hitz prozesatzen du leiho irristakor baten bidez, eta behatutako hitza (lauki urdina) bere testuinguruko hitzak (lauki zuria) aurresateko erabiltzen du. Adibide horretan esandakoari jarraiki, (3.1) ekuazioak Skip-gram ereduak maximizatzen duen probabilitatea deskribatzen du:

[4] `https://code.google.com/archive/p/word2vec/`

$$argmax\frac{1}{T}\sum_{t=1}^{T}\sum_{-k<j<k,j!=0} log(p(c_{t+j}|w_t) \tag{3.1}$$

Ekuazio horretan w_t behatutako hitza da, c_{t+j} testuinguruko hitzak (j bidez indexatuak), k testuinguru-leihoaren tamaina, T corpuseko hitz kopurua eta p maximizatu beharreko hitz-testuinguru probabilitatea. Azken horrek k testuinguru-leihoaren baitako (w_t,c_{t+j}) agerkidetzak behatzeko probabilitatea adierazten duenez, p maximizatzeko Skip-gram ereduak w_t eta c_{t+j} optimizatzen ditu testu-corpuseko agerkidetza guztietan. CBOW ereduan metodo bera erabiltzen da, baina sarreran testuinguruko c_{t+j} hitzak izango ditugu.

Softmax funtzioa neurona-sarren irteran jartzen da, eta ataza jakin bateko klase anitzeko probabilitateak kalkulatzen ditu emaitza legez. (3.1) ekuazioko nomenklatura berari jarraiki, (3.2) ekuazioak behatutako hitz-testuinguru pare batentzako *softmax* funtzioa deskribatzen du[5]:

$$p_{smax}(c_{t+j}|w_t) = \frac{e^{c_{t+j}w_t}}{\sum_{i=1}^{V} e^{c_i w_t}} \tag{3.2}$$

Ekuazio horretan w_t behatutako hitzaren bektorea da, c_{t+j} behatutako testuinguruarena, c_i behatutakoa ezik beste edozein testuingururen bektorea, eta V corpuseko hiztegiaren tamaina. Kontuan izan, *softmax* funtzioaren izendatzailea oso garestia dela konputazionalki, V operazio[6] egin behar baitira. Bada, laginketa negatiboan kalkulu garesti hori asko arintzen da, eta, (w, c_i) guztiak erabili beharrean, N testuinguru ausaz aukeratzen dira. N txikia izaten bada ere, laginketa negatiboa *softmaxaren* hurbilpen eraginkorra da (Goldberg and Levy, 2014) eta, ordura arteko neurona-sare ereduekin alderatuta, ikasketa-prozesua asko azkartu du (Mikolov *et al.*, 2013a, c). Horixe da, hain zuzen, laginketa negatiboa `word2vec` eredu-multzoaren arrakastaren gakoetako bat, *softmax* funtzioaren hurbilpen eraginkorra bezain azkarra baita.

Esanak esan, hurrengo paragrafoetan Skip-gram ereduko galera-funtzioaren eta laginketa negatiboaren xehetasuna deskribatuko ditugu. Bada, Skip-gram ereduak bi hitz-bektore esleitzen dizkio hiztegiko sarrera bakoitzari,

[5]Kasu honetan, klaseak testu-corpuseko hiztegiko sarrea guztiak izango dira.

[6]Behatutakoa k parearekin ezik, beste c_i guztiekin $e^{c_i^t w}$ kalkulatu behar da. Hau da, V kalkulu egin behar dira.

hain zuzen, hitz eta testuinguru moduan dituzten tasun semantikoak adierazteko. Lan honetan hitz eta testuinguru matrizeak W eta C legez izendatuko ditugu, eta ikasi beharreko parametroak dira. Bi horiek $|V| \times D$ tamainako matrizeak dira, $|V|$ hiztegiaren tamaina izanik eta D hitz-bektoreen dimentsionaltasuna. Ereduaren inplementazioari jarraiki, Skip-gram ereduko deskribapenean C eta W izendatutako matrizeak `word2vec` eredu-multzoko sarrerako eta irteerako pisuen matrizeak dira, hurrenez hurren.[7]

(3.3) ekuazioak hitz-testuinguru agerkidetza bakoitzerako $J_{sg} : W \subset \mathbb{R}^D \times C \subset \mathbb{R}^D \to \mathbb{R}$ galera-funtzioa definitzen du, laginketa negatiboa barne. Hala, $w \in W$ behatutako hitza da, $c \in C$ testuinguru-hitza, $w_n \in C$ ausaz $P(c)$ zarata-distribuziotik erauzitako hitza (n bidez indexatua), eta σ sigmoide funtzioa.

$$J_{sg}(w, c) = log(\sigma(w^T c)) + \sum_{n=1}^{N} \mathbb{E}_{w_n \sim P(w)} \Big[log(\sigma(-w_n^T c)) \Big] \tag{3.3}$$

Galera-funtzio horretan agertzen ez bada ere, azpi-laginketa atalasea (*sub-sampling threshold*) deituriko parametroak maiztasun handiko hitzak ikaskuta-prozesutik ezabatzen ditu. Parametro hori (3.4) ekuazioak azaltzen du:

$$p = 1 - \sqrt{\frac{t}{f}} \tag{3.4}$$

Aipatutako ekuazioan f hitzak corpusean duen maiztasuna da, t atalasea, eta p behatutako hitza ezabatzeko probabilitatea. Algoritmoak hitz-bektoreak kalkulatzen hasi aurretik, f maiztasuna t atalasea baino altuagoa duten hitzak p probabilitatearekin ausaz ezabatzen ditu.

Galera totala kalkulatze aldera, corpuseko hitz-testuinguru agerkidetza guztietarako (3.3) ekuazioa kalkulatu beharko genuke, eta, ondoren, denak batu (betiere K tamainako testuinguru-leihoaren baitan). Ereduaren hiper-parametro nagusiak K testuinguru-leihoa, N lagin negatibo kopurua, t atalasea eta dimentsionaltasuna dira. Bada, `word2vec` eredu-multzoak (3.3)

[7] Hainbat autorek `word2vec` kodeko sarrerako pisuen matrizea "hitzen" matrize legez definitzen dute, baina, kodearen inplementazioari so, sarrerako pisuen matrizea testuinguruko hitzen bektoreak dituena da (bai Skip-gramen eta bai CBOWen). Inplementazioa oinarri legez hartuta, eta nahasmena saiheste aldera, lan honetan hurrengo irizpidea jarraituko dugu: kodeko sarrerako matrizea gure lanean testuinguruen C matrizea da, eta kodeko irteerako matrizea W hitzen matrizea. A.1 eranskinean Skip-gram ereduko kodezati bat azaltzen ditugu; besteak beste, laginketa negatiboa eta aipatutako matrizeen eguneraketak deskribatzen ditugu.

ekuazioa gradientea-jaitsiera estokastikoaren bitartez optimizatzen du, eta, modu horretan, corpus jakin baten galera-funtzioa maximizatzeko W eta C optimoak kalkulatzen ditu. C ausaz hasieratzen da eta W zerora. A.1 eranskinean Skip-gram ereduko kode-zati bat deskribatzen dugu; besteak beste, galera-funtzioaren gradienteak eta W eta C matrizeen eguneraketak.

Skip-gram ereduarekin ikasitako hitz-bektoreek sintaxia eta semantika ondo modelatzeko gaitasuna erakutsi dute. C matrizeko hitz-bektoreak hitzen arteko analogia eta antzekotasuna erreproduzitzeko erabiltzen dira, hitz-bektoreen kosinua medio (Baroni *et al.*, 2014). Izan ere, gizakiek osatutako antzekotasun eta ahaidetasun urre-patroiekin burututako ebaluazioek C matrizeko hitz-bektoreen kalitatea baieztatu dute (Baroni *et al.*, 2014). Horrexegatik, hain zuzen, 3. eta 4. kapituluko ebaluazio oro C matrizeko hitz-bektoreen gainean egin dira. 5. kapituluan W eta C matrizeen informazio osagarria aztertuko dugu.

3.2 Esperimentuak

Atal honetan guk proposatutako ausazko ibilbideen metodoaren inguruko esperimentuen nondik norakoak deskribatuko ditugu. Lehenik, esperimentu horietan erabilitako baliabideak aurkeztuko ditugu, ondoren, emaitzak, eta, azkenik, emaitzetatik ateratako ondorioak.

3.2.1 Baliabideak

Kapitulu honetako esperimentuetako metodoak eta baliabideak 3.3. taulan laburtu ditugu. Gure baliabide motak hurrengoak dira: datu-multzoak, corpusak eta ezagutza-base bat. Hitzen errepresentazioei dagokienez, lau metodo bereizten ditugu. Baliabide eta metodo guztien laburdurak hurrengo paragrafoetan azaltzen joango dira, eta kapitulu honetako hitzen errepresentazioak metodo-baliabide hitz-konbinaketa batekin adieraziko ditugu.

Erabilitako bektore oro hurrengo bi datu-multzoetan ebaluatu ditugu: WordSim353 (WS) (Finkelstein *et al.*, 2001) eta SimLex999 (SL) (Hill *et al.*, 2015). Aurreneko datu-multzoan ahaidetasuna eta antzekotasuna nahastuta daude eta bigarrena antzekotasunekoa da bakarrik. Gogoan hartu Spearmanekin giza irizpideen eta sistemaren emaitzen sailkapen-balioen arteko korrelazioa kalkulatzen dela.

		Laburdura	Deskribapena
Baliabideak	Datu-multzoa	WS	WordSimS353 (ahaidetasuna)
		SL	SimLex999 (antzekotasuna)
	Corpus mota	wn	Wordnet corpus sintetikoa (edo grafoa)
		gnw	Google News testu-corpusa
	Ezagutza-baseak	wnEN	Ingelesezko WordNet
Metodoak		PPB	Personalized PageRank bektoreak
		CB	CBOW testu-corpusetan
		AICB	CBOW corpus sintetikoetan
		AISG	Skip-gram corpus sintetikoetan

3.3 taula – Kapitulu honetako baliabideen laburdurak eta deskribapenak. Goiko bi lerroek esperimentuotan erabilitako datu-multzoen laburdurak azaltzen dituzte (WS eta SL); hurrengo bi lerroek corpus motak (*gnw* eta *wn*), eta, ondorengoak, erabilitako ezagutza-base bakarra (*wnEN*). Azkenengo lau lerroetan hitzen errepresentazioak sortzeko metodoak agertzen dira; hots, Personalized PageRank bektoreak PPB legez, CBOW testu-corpusetan aplikatzean CB moduan, CBOW corpus sintetikoetan aplikatzean AICB moduan eta Skip-gram corpus sintetikoetan aplikatzean AISG moduan. Azkenengo bi metodoekin kalkulatutako errepresentazio trinkoak eta *wn* corpus sintetikoa berriak dira, lan honetan proposatutakoak.

Oinarri-lerroko sistema moduan, ezagutza-baseetan eta testuetan oinarritutako metodoak erabili ditugu. Ezagutza baseetan oinarritutako bektoreei dagokienez, UKBrekin[8] glosadun WordNet 3.0 (*wnEN*) erabili dugu grafo moduan, eta WordNetek berak dakarren hiztegia. WordNet bertsio horrek 117.659 erpin eta 525.356 ertz ditu eta aipatutako hiztegiak hitzak eta *synsetak* lotzen ditu; zehazki, WordNeten egileek eskuz anotatutako corpus batean oinarrituta, hitz bakoitzaren adieren (*synseten*) maiztasunak adierazten dituzte. Bai grafoa eta bai hiztegia eskuragarri daude[9]. Azken bi horietatik abiatuta, UKB bidez Wordnetetik erauzitako Personalized PageRank bektoreak (PPB) (Agirre and Soroa, 2009; Agirre *et al.*, 2014) kalkulatu ditugu, $\alpha = 0,85$ balio lehenetsiarekin (Agirre *et al.*, 2010).

Kontuan izan, Personalized PageRank ausazko ibilbideetan oinarritutako metodoa dela (ik. 2.2. atala), eta guk proposatutako ausazko ibilbideen metodoa azken horretan oinarritzen dela. Gauzak horrela, guk proposatutako ausazko ibilbideen metodoa ere UKB gainean garatu dugu. UKB moldatu

[8] `http://ixa2.si.ehu.es/ukb/`
[9] `http://ixa2.si.ehu.es/ukb/`

egin dugu grafo gainean eginiko ausazko ibilbideetan kontzeptuen lexikalizazioak fitxategi batetara emititu ditzan. PPBekin bezala, ausazko ibilbide horiek glosadun WordNet 3.0 gainean aplikatu ditugu, $\alpha = 0,85$ balio lehenetsiarekin (Agirre *et al.*, 2010), eta I testuinguru kopurua 70 milioirekin. Ausazko ibilbideak gauzatzeko parametroak, beraz, defektuzkoak dira, eta, hurrengo atalean ikusiko dugun bezala, corpus tamainarekin soilik esperimentatu dugu.

WordNeteko testuinguru sintetikoekin osatutako corpus sintetikoa `word2vec` (Mikolov *et al.*, 2013a) eredu-multzoarekin prozesatu dugu, azken horren bi arkitektura posibleekin; ausazko ibilbideekin osatutako corpusa CBOWrekin prozesatzeari AICB deituko diogu eta Skip-gramekin AISG. Gainera, aipatutako bi eredu horietan parametro berak erabili ditugu: 300eko dimentsionaltasuna, 3 iterazio, 5 lagin negatibo eta 5eko leihoa. Dimentsionaltasuna izan ezik, balio horiek `word2vec` eredu-multzoaren parametro lehenetsiak dira.

Testu hutseko hitz-bektoreei dagokienez, `word2vec` eredu-multzoaren CBOW (Mikolov *et al.*, 2013a) ereduarekin (CB) aurre-entrenatutako hitz-bektoreak[10] erabili ditugu, $100 \cdot 10^9$ tokeneko *Google News* corpusaren (*gnw*) gainean kalkulatukoak . Autoreen aburuz, hurrengo parametroekin kalkulatu dira hitz-bektoreak: 300 dimentsio, 3 lagin negatibo, 10^{-5} azpi-laginketa atalasea eta 5eko leiho zabalera.

Hitzen errepresentazioen metodo-baliabide notazioa sinplifikatze aldera, eta hiru metodoek WordNet baliabidea darabiltela buruan, AISG, AICB eta PPB dauden metodo-baliabide konbinaketetan *wn* erabiliko dugu corpus sintetikoak eta grafoa izendatzeko[11]. Gauzak horrela, kapitulu honetan metodo-baliabide konbinaketa posibleak hurrengoak dira: PPB_{wn}, CB_{gnw}, AICB_{wn} eta AISG_{wn}. Aurreneko biak gure oinarri-lerroak dira, eta, azkenak, gure proposamenak. Horiek guztiak zehaztu ondoren, hurrengo atalean esperimentuen xehetasunak eta lortutako emaitzak azalduko ditugu.

3.2.2 Emaitzak

Ebaluazioa gauzatze aldera, hitzen antzekotasunaren eta ahaidetasunaren inguruko literatura mardula jarraituko dugu (Agirre *et al.*, 2009a) 3., 4. eta 5. kapituluetan. Esperimentu guztietako bektore mota guztiekin antzekotasun-balioak kalkulatzeko hitzen errepresentazioen arteko kosinua erabili dugu

[10] `https://code.google.com/archive/p/word2vec/`

[11] 4. eta 5. kapituluan ere irizpide bera erabiliko dugu.

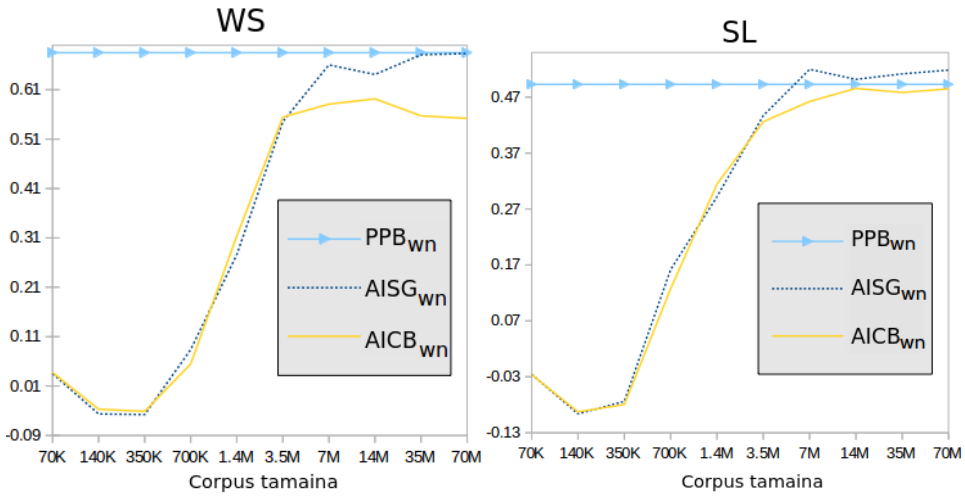
(ik. 1.2. atala). Salbuespen bakarrak bi izan dira; 4. kapituluko bektoreen konbinaketa batean bektore konplexuan antzekotasuna erabili dugu (Scharnhorst, 2001), eta 5. kapituluko NASARI bektoreekin *weighted overlap* metodoa (Pilehvar *et al.*, 2013). Horrez gain, ohizko praktikari jarraiki (Ruder *et al.*, 2017), lan honetako bektore guztien erantzunak ebaluatzeko urre-patroien eta metodoen erantzunen arteko Spearman (Spearman, 1904) korrelazioaren bitartez egin dugu. Atal honetako emaitzak hurrengo urratsetan azalduko ditugu: lehenengoan, corpus sintetikoaren tamainak ebaluazioan duen eragina neurtuko dugu, eta, bigarrenean, testu eta ezagutza-baseen informazio semantikoa konbinatzeko teknika soila burutuko dugu.

Corpus sintetikoaren tamainaren eragina aztertzen

Esperimentu hauetan 3.3 atalean proposatutako metodoaren bi aldaera erabili ditugu. Bada, testuinguru sintetikoekin corpusa osatu ondoren, Skip-gramekin (AISG_{wn}) edo CBOWekin (AICB_{wn}) azken horren hitz-errepresentazio trinkoak kalkulatzen ditugu. Bada, testuinguru sintetikoen kopurua handitu ahala, antzekotasun datu-multzoen gainean AISG_{wn} eta AICB_{wn} metodoetako bektoreen emaitzak ikusi ditugu lehenik.

Gauzak horrela, 3.3(a) eta 3.3(b) aipatutako metodoek WS eta SL datu-multzoen gaineko ikasketa-kurbak erakusten dituzte, hurrenez hurren.

Bi datu-multzoetako kurbei so, AISG_{wn} eta AICB_{wn} metodoen bektore-adierazpenen kalitatea azkar hobetzen doa $7 \cdot 10^6$ testuingurura arte, eta $70 \cdot 10^6$ testuingurura heltzerakoan emaitzak konbergitzen hasten dira. Konbergentzia horren balioak 3.4. taulan azaltzen dira, PPB_{wn} eta CB_{gnw} bektoreenekin batera. Bada, WS datu-multzoan AICB_{wn}ek ezik, beste bektoreek antzeko emaitzak lortzen dituzte. SLeri dagokionez, AISG_{wn}en emaitzak gailentzen dira. Emaitzok ikusita, gure metodoek WordNeteko informazioa modu eraginkorrean jasotzeko gai direla ondorioztatzen dugu, ezagutza-baseetan oinarritutako PPB_{wn}en pareko emaitzak (WSen) edo hobeak (SLen) lortzen baitituzte. Kontuan izan, momentu hartan WSen WordNet erabilita izandako emaitzarik hoberena Agirre *et al.*-ek (2010) izan zuela; hots, PPB_{wn}en bidez 68,5 Spearman balioa. Horrez gain, testu-corpusetan oinarritutako CB_{gnw}ren pare dago WSen eta nahiko hobeto SLen.



(a) WS: Ahaidetasuna **(b)** SL: Antzekotasuna

3.3 irudia – Spearman balioak WS (a) eta SL (b) urre-patroietan, PPB_{wn}ekin eta proposatutako AISG_{wn} eta AICB_{wn} metodoekin. Ardatz bertikalean Spearman balioak, eta horizontalean testuinguru sintetikoaren tamaina. PPB_{wn}ekin adierazitako balioa da oinarri-lerroa (ez dago corpus tamainaren menpe). AISG_{wn} eta AICB_{wn} metodoentzat, testuinguru sintetiko kopurua handitu ahala Spearman balioek duten eboluzioa azaltzen da.

Testuaren eta WordNeten informazioa konbinatzen

Azken horiek ikusita, eta metodo desberdinen osagarritasuna aztertze aldera, gure metodorik hoberena (AISG_{wn}) PPB_{wn}ekin eta CB_{gnw}ekin konbinatu dugu, eta, horretarako, hiru metodo horien emaitzez baliatuko gara. Erabilitako bektoreen arabera, antzekotasun eskalak oso desberdinak izan daitezkeela kontuan izanik, espazio bereizietako datu-multzoetako hitz-pareen antzekotasun-emaitzak txikienetik handienera ordenatu ditugu, eta, ondoren, pare bakoitzari espazio bereizietan erdietsitako sailkapen-balioen batezbestekoa esleitu diogu (RNK laburdura). 3.5. taularen lehen hiru lerroetan aipatutako hiru metodoekin WS eta SL datu-multzoetan lortutako emaitzak erakusten dira, erdialdeko hiru lerroetan metodoen konbinaketa posible oro, eta, azkenengoan, orduko artearen egoera.

Bada, RNK(ab) konbinaketa bi metodoak bereizita baino hobea da bi datu-multzoetan, eta, AISG_{wn} eta PPB_{wn} metodoetako informazioa desberdi-

	SL	WS
CB_{gnw}	44,2	**68,6**
PPB_{wn}	49,3	68,3
AISG_{wn}	**52,0**	68,3
AICB_{wn}	48,6	59,1

3.4 taula – Spearman balioak SL eta WS urre-patroietan, zutabeetan. Lehenengo lerroan testu-corpusetan oinarritutako aurre-entrenatutako bektoreekin lortutako emaitzak (CB_{gnw}), eta bigarrenean UKB WordNet gainean aplikatuta lortutako bektoreenak (PPB_{wn}). Azken bi lerroetan guk proposatutako metodoen emaitzak: WordNet gainean ausazko ibilbideak aplikatuta, lortutako testuinguru sintetikoak Skip-gramekin (AISG_{wn}) eta CBOWekin (AICB_{wn}) prozesatzerakoan. Emaitzarik hoberenak urre-patroi bakoitzean letra lodiz.

na baina osagarria dela ondorioztatzen dugu. RNK(ac) konbinaketan ere bi metodoak bereizita baino hobea da bi datu-multzoetan, WordNeten (AISG_{wn}) eta testuan (CB_{gnw}) dagoen informazioaren osagarria dela ondorioztatuz. Kontuan izan RNK(ab) konbinaketarekin alderatuta, SLeko emaitzak berdintsuak badira ere, WS datu-multzoan *ac* konbinaketarekin (WordNeteko eta testu hutseko informazioaren konbinaketarekin) emaitzak ia bost puntu hobetzen direla. RNK(abc) lerroan fenomeno bera ikusten da, eta emaitzak are gehiago hobetzen dira bi datu-multzoetan, bai metodo bereiziekiko eta bai aurreko bi konbinaketekiko. Are gehiago, hiru metodoen konbinaketarekin SL datu-multzoan momentu hartan publikatutako emaitzarik hoberena lortu genuen, eta orduko artearen egoeratik hurbil geratu ginen WSen. Bada, garai horretako 3.5 taulako SLen artearen egoera Hill *et al.*-ek (2014) zeukan, eta corpus elebidun baten gainean entrenatutako neurona-sare errekurrente baten bidez lortu zuen. Taula bereko WSen artearen egoera Radinsky *et al.* (2011) lanari dagokio, zeinek Wikipedian oinarritutako algoritmo bat testuan oinarritutako batekin konbinatu zuen.

Kontuan izan guk konbinaketa erraz batzuk burutu ditugula informazio iturri bereizien osagarritasuna aztertzeko. Konbinaketa sofistikatuagoek are gehiago hobetu ditzakete emaitzak. Hurrengo atalean testu-corpusetako eta ezagutza-baseetako informazio semantikoa uztartzeko hainbat konbinaketa proposatuko ditugu.

Analisi kualitatibo batzuk egin ondoren, hurrengoak aurkitu ditugu: al-

	SL	WS
(a) AISG_{wn}	52,0	68,3
(b) PPB_{wn}	49,3	68,3
(c) CB_{gnw}	44,2	68,6
RNK(ab)	53,5	70,0
RNK(ac)	53,3	74,8
RNK(abc)	**55,2**	75,9
Artearen egoera	52,0	**80,0**

3.5 taula – SL eta WS datu-multzoetako Spearman balioak zutabeetan. Lehenengo hiru lerroetan 3.4 taulan aipatutako hiru metodoren emaitzak. Metodo horietako bakoitzari letra bat esleitzen zaio: *a* letra AISG_{wn} da, *b* letra PPB_{wn} eta *c* letra CB_{gnw}. Hurrengo hiru lerroetan aipatutako metodoen RNK konbinaketarekin lortutako emaitzak. Azkenengo lerroan, artearen egoeran zeuden emaitzak. Emaitzarik hoberenak urre-patroi bakoitzean letra lodiz.

de batetik, testu-corpusetan oinarritutako hitz-bektoreek (esperimentuetan erabilitako leiho zabalera berarekin) hitz ahaideak multzokatzeko joera arina dutela[12], eta, hein nahiko txikiagoan hitz antzekoak multzokatzekoa; beste aldetik, ezagutza-baseetan oinarritutako gure proposameneko errepresentazio trinkoek hitz ahaideak zein antzekoak multzokatzen badituzte ere, antzekotasunerako joera argia erakusten dute.

Analisi hori bat dator 3.4. taulako emaitzekin, AISG_{wn} eta AICB_{wn} CB_{gnw} baino hobeto baitabiltza SL antzekotasun datu-multzoan. 4. eta 5. kapituluetan antzekotasun-emaitzak hobetzeko joera bera agertzen da, WordNet bidez sortutako errepresentazio trinkoek antzekotasun hutseko datu-multzoetako emaitzek hobekuntza nabarmenak baitituzte (ik. 4.2.2. eta 5.3.2. atalak).

3.3 Ondorioak

Atal honetan proposatutako algoritmoak WordNeteko egitura bektore-espazio jarrai batean kodetzen du. Horretarako, grafoen gaineko ausazko ibilbideak hizkuntza-ereduetan oinarritutako neurona-sare batekin prozesatu dugu, eta

[12] Esaterako, *physics-proton*

errepresentazio trinko mota berriak sortu ditugu. Antzekotasun eta ahaidetasun datu-multzoetan eginiko ebaluazioek gure hitzen errepresentazio trinko berriek jatorrizko Personilized PageRank bektoreen (PPB_{wn}) emaitzak berdintzen edo hobetzen dituztela erakusten dute, baina, hamarka milako dimentsio erabili beharrean 300ekin soilik. Horrez gain, corpus sintetikoetatik Skip-gramekin edo CBOWekin kalkulatutako errepresentazio trinkoek ((AISG_{wn} eta AICB_{wn}, hurrenez hurren) Personilized PageRank bektoreen (PPB_{wn}) eta testutik erauzitako hitz-bektoreen (CB_{gnw}) osagarriak direla erakutsi dugu; azken horien emaitzak konbinatuta hobekuntza nabarmenak erdietsi ditugu, eta lan hura publikatu zen garaiko SimLex999 (SL) antzekotasun datu-multzoko artearen egoera gainditu dugu.

4

Testu eta ezagutza-baseen konbinaketa

Aurreko kapituluan aurkeztutako ezagutza-baseetan oinarritutako metodoa oinarri legez hartuta, kapitulu honetan testu eta ezagutza-baseetako informazio osagarria uztartzeko konbinaketetan sakonduko dugu. 2. kapituluan ikerketa-ildo bereko hainbat metodo deskribatu ditugu (Faruqui and Dyer, 2014; Halawi *et al.*, 2012; Rastogi *et al.*, 2015; Tian *et al.*, 2015), denak ere estrategia berarekin; testu bektore-espazioa ezagutza-baseetako erlazioen murriztapenekin aberastea. Gure proposamena, ordea, beste ikuspegi batean oinarritzen da, testu eta ezagutza-baseetako espazio bereiziak konbinatzea proposatzen baitugu. Izan ere, bektore-espazioek ezagutza-baseko informazio estrukturala modu oso eraginkorrean kodetzen dutela ikusita, 2. kapituluko metodoen murriztapenak baino informazio semantiko aberatsagoa duten hipotesitik abiatu gara. Gainera, testutik eta ezagutza-baseetatik erauzitako errepresentazio trinkoak eta corpus sintetikoek bi espazioen arteko hainbat konbinaketa ahalbidetzen dituzte, bi iturri bereizietako informazioen osagarritasuna ondo ustiatzeko gai direnak.

Guk proposatutako konbinaketak hitzen antzekotasun eta ahaidetasun datu-multzoetan ebaluatu ditugu, momentu hartako testu-corpusak eta ezagutza-baseak uztartzeko teknika batzuk tarte handiagatik gaindituz.

4.1 Metodoa

Atal honetan izaera desberdineko hitzen errepresentazioen konbinaketa deskribatuko ditugu. Sarreran aipatu bezala, errepresentazio horiek testu hu-

		Laburdura	Deskribapena
Konbinaketak	Bektoreak	KAT	Bektoreen arteko kateaketa
		ZEN	Bektoreen arteko zentroidea
		KNP	Zenbaki konplexuak osatu bektoreekin
	Korrelazio bidez	KKA	Bektore-espazioen arteko korrelazio kanonikoaren analisia
		ONA	Kateatutako bektoreen osagai nagusien analisia
	Corpusak	HIB	Testuaren eta corpus sintetikoaren konbinaketa
	Emaitzak	BB	Emaitzen batezbestekoa
		RNK	Emaitzen sailkapen-balioetan oinarritua

4.1 taula – Kapitulu honetako konbinaketen laburdurak eta deskribapenak. Lau taldetan banatzen dira: bektoreen arteko konbinaketak (KAT, ZEN eta KNP), bektoreen arteko korrelazioenak (KKA eta ONA), corpusen konbinaketak (HIB) eta emaitzen konbinaketak (BB eta RNK). Konbinaketa horiek guztiak testu-corpusetako eta ezagutza-baseetako informazio semantikoa uztartzeko proposamenak dira.

tsean eta ezagutza-baseetan oinarritutakoak dira, eta hainbat iturri bereizi konbinatu ditugu, formatu desberdinetan. Behar izan dugunean, RG datumultzoaz (Rubenstein and Goodenough, 1965) baliatu gara (garapen datumultzo legez) parametroak optimizatzeko. Bada, guk proposatutako konbinaketak eta beren laburdurak 4.1. taulan daude laburbilduta. Taula horretan ikusten denez, lau taldetan sailkatu ditugu konbinaketak: (1) bektoreena, (2) korrelazio bidezkoa, (3) corpusena eta (4) emaitzena. Errepresentazio horiek guztiak hibrido legez izendatuko ditugu.

4.1.1 Bektoreen konbinaketa

Testu hutseko eta ezagutza-baseetako errepresentazio trinkoak modu oso errazean konbinatu daitezke, hala nola, kateaketaren (KAT), zentroidearen (ZEN), edo zenbaki konplexuen (KNP) bidez. Azken hori Wittek *et al.*-en (2013) proposamenean oinarrituta dago; hots, bektore konplexuaren osagai erreala semantika distribuzionala errepresentatzeko erabiltzea, eta, osagai konplexua ontologiaren informazioa kodetzeko. Proposamen horri jarraiki, guk testuan oinarritutako hitz-bektoreak osagai errealean sartu ditugu, eta WordNeten oinarritutako errepresentazio trinkoak konplexuan.

Aipatutako hiru konbinaketak dimentsionaltasun bereko bektoreetan soilik aplikatu daitezke. Kateaketaren kasuan, posible bada ere, dimentsionaltasun handiena duenak gailentzeko joera izango luke antzekotasun emaitzetan.

4.1.2 Korrelazio bidezko konbinaketa

Osagai nagusien analisia (ONA) (Hotelling, 1933) dimentsionaltasuna murrizteko teknika da. Bada, gure esperimentuetako datu-multzoetan (ik. 4.2. atala) agertutako hitzen errepresentazio kateatuei ONA aplikatu diegu. ONAren dimentsionaltasuna RG datu-multzoan optimizatu dugu, eta emaitzarik hoberenak 300ekin lortu ditugu. ONArekin, beraz, KATen informazioa dimentsionaltasun murriztuagoarekin errepresentatzen da.

Korrelazio kanonikoaren analisiak (KKA) ere korrelazio linealak ditu aztergai, baina, kasu honetan, bi aldagai multidimentsionaleko multzoen artekoak. Jatorrian, KKA ele desberdinetako bektore-espazio bereiziak espazio komun batean proiektatzeko erabiltzen da (Faruqui and Dyer, 2014). Guk, baina, hitz bereko testu eta WordNeteko errepresentazio trinkoak espazio komun batera proiektatuko ditugu, non bi horien arteko korrelazioa maximizatuko den. Espazio komun hori sortzerakoan bi aukera daude: alde batetik, dimentsionaltasuna optimizatzea, eta, bestetik, testu bektore-espazioa edo ezagutza-basekoa espazio komunera proiektzatzea. Horiak guztiak RG datu-multzoaren gainean probatu ditugu, eta 180 dimentsioko WordNeteko errepresentazio trinkoak espazio komunean proiektatzerakoan lortu ditugu emaitzarik hoberenak.

4.1.3 Corpusen konbinaketa

Testuan oinarritutako errepresentazioek eta Wordnetekoek izaera semantiko desberdinetako corpusak erabiltzen dituzte. Iturri semantiko desberdin horietako informazioa uztartze aldera, lehenik, corpus hibrido batean konbinatu ditugu, eta, ondoren, Skip-gram bidez prozesatu eta hitz-bektore hibridoak kalkulatu (HIB). Ikerketaren momentu honetan ez dugu konbinaketan parte hartzen duten faktorerik aztertu; besterik gabe, momentuko testu eta WordNet corpusen lerroak nahastu, eta Skip-gram bidez hitzen errepresentazioak kalkulatu ditugu. Azken horren parametroak 4.2.1. ataleko WBU corpusa prozesatzeko erabilitako berak dira, ez dugu optimizatu.

Corpus hibrido elebakar horren osaketan, testu eta WordNet corpusetako token kopuruen kontrola faktore garrantzitsua da (ik. 5.2.3. atala, hurrengo kapituluan), bi espazio horietako informazioen osagarritasuna hobeto ustiatzen baitu. Hala ere, ikerketaren fase honetan ez dugu halakorik esploratu, konbinaketaren eraginkortasunean zentratu gara soilik.

4.1.4 Emaitzen konbinaketa

Metodo honek datu-multzoetako hitz-pare bakoitzeko antzekotasun- edo ahaidetasun-balioak konbinatzen ditu; hots, testuan eta WordNeten oinarritutako metodoekin lortutako emaitzak. Bada, bi espazio horietako hitzen errepresentazioen kosinuen balioen batezbestekoa (BB) kalkulatuko dugu. Azken horren aldaera legez, eta 3.2.1. atalean egindakoari jarraiki, bi espazioetako datu-multzoak antzekotasun- edo ahaidetasun-balioen arabera (txikienetik handienera) ordenatu ditugu, eta, ondoren, sailkapen-balioen batezbestekoa kalkulatu dugu (RNK). 3.3. atalean aipatu bezala, sailkapen-balioak konbinatzearen motibazioa Spearmanek, antzekotasun- edo ahaidetasun-balioak kontuan hartu barik, korrelazioa sailkapen-balioetan oinarrituta dagoela da. Gogoan izan atal BB eta RNK metodoetan ez dela hitzen errepresentaziorik ez corpusik konbinatzen.

4.2 Esperimentuak: bi iturri konbinatzen

Kapitulu honetan erabilitako baliabide eta metodo oro 4.2. taulan laburbildu ditugu. Aurreko kapituluan erabilitako baliabide eta metodoak beretsuak dira, baina, anitzagoak. Baliabide eta metodo guztien laburdurak hurrengo ataletan azaltzen joango dira. Lan osoan egingo dugun legez, hitzen errepresentazioen arteko antzekotasuna kosinuaren bidez kalkulatuko dugu. Kapitulu honetako salbuespen bakarra KNP da, kasu horretan bektore konplexuen arteko antzekotasuna erabili baitugu (Scharnhorst, 2001).

4.2.1 Baliabideak

Kapitulu honetako esperimentuetan testu-corpusetatik eta ezagutza-baseetatik eratorritako informazioa ikasten dugu, eta, behar izanez gero, parametroak RG antzekotasun datu-multzo optimizatu ditugu.

Testuetan oinarritutako errepresentazioak

Baroni *et al.* (2014) lanari jarraiki, ingelesezko Wikipedia[1], *Brittish National Corpus*[2] eta *ukWaC*[3] kateatuta testu-corpus bat sortu dugu, WBU

[1] `linguatools.org/tools/corpora/wikipedia-monolingual-corpora/`
[2] `http://www.natcorp.ox.ac.uk`
[3] `http://wacky.sslmit.unibo.it`

		Laburdura	Deskribapena
Baliabideak	Datu-multzoa	WS	WordSimS353 (ahaiadetasuna)
		MEN	MEN (ahaiadetasuna)
		MTU	MTURK287 (ahaiadetasuna)
		WSR	WordSimS353 Relatedness (ahaiadetasuna)
		WSS	WordSimS353 Similarity (antzekotasuna)
		RG	RG (antzekotasuna)
		SL	SimLex999 (antzekotasuna)
	Corpus mota	wn	Wordnet corpus sintetikoa (edo grafoa)
		wiki	Wikipedia corpus sintetikoa (edo grafoa)
		wbu	WBU testu-corpusa
		gnw	Google News testu-corpusa
	Ezagutza-baseak	wikiEN	Ingelesezko Wikipedia
		wnEN	Ingelesezko WordNet
Metodoak		PPB	Personalized PageRank bektoreak
		CB	CBOW testu-corpusetan
		SG	Skip-gram testu-corpusetan
		AISG	Skip-gram corpus sintetikoetan

4.2 taula – Kapitulu honetako baliabideen laburdurak eta deskribapenak. Goiko zazpi lerroek esperimentuotan erabilitako datu-multzoen laburdurak azaltzen dituzte, aurreneko laurak ahaidetasunezkoak (WS, MEN, MTU eta WSR), eta hurrengo hirurak antzekotasunezkoak (WSS, RG eta SL). Hurrengo lau lerroek corpus motak adierazten dituzte, bi corpus sintetiko (*wn* eta *wiki*), eta, beste bi testu hutsekoak (*wbu* eta *gnw*). Ondorengo lerroek erabilitako bi ezagutza-baseak azaltzen dituzte (*wnEN* eta *wikiEN*). Azkenengo lau lerroetan hitzen errepresentazioak sortzeko metodoak agertzen dira; hots, Personalized PageRank bektoreak PPB legez, CBOW testu-corpusetan aplikatzean CB moduan, Skip-gram testu-corpusetan aplikatzean SG moduan, eta Skip-gram corpus sintetikoetan aplikatzean AISG moduan. Azkenengo metodoarekin kalkulatutako errepresentazio trinkoak eta *wn* eta *wiki* corpusak berriak dira, lan honetan proposatutakoak.

deitu duguna (*wbu*). Corpus hori $5 \cdot 10^9$ tokenez osatuta dago, eta Skip-gram (Mikolov *et al.*, 2013a) (SG) aplikatuta bere hitz-bektoreak kalkulatu ditugu. Bada, parametroak RG datu-multzoaren gainean optimizatu ditugu, eta hurrengoekin lortu ditugu emaitzarik hoberenak: 300eko dimentsionaltasuna, 5 lagin negatibo, azpi-laginketa atalasea 0, eta leiho zabalera 5.

WordNeten oinarritutakoak errepresentazioak

Aurreko kapituluan deskribatutako ausazko ibilbideen metodoa erabili dugu, ezagutza-baseak eta semantika distribuzionala konbinatuta WordNeten egitura bektore-espazio batean kodetzen duena. Izan ere, WordNeteko informazioa bektore-espazio batean egoterakoan, dimentsionaltasun bereko testu hitz-bektoreekin aise konbinatu daiteke.

Bada, aurreko kapituluan deskribatutako metodorik hoberena erabili dugu, hots, corpus sintetikoen gainean Skip-gram aplikatzea (AISG). Beraz, lehenik UKBren bitartez glosadun WordNet 3.0 gainean (*wnEN*) ausazko ibilbideak burutu ditugu, 3. kapituluko ibilbideen $\alpha = 0,85$ balio lehenetsia (Agirre and Soroa, 2009) erabilita. AISGn egindakoari jarraiki, aipatutako WordNet corpusa Skip-gram ereduarekin prozesatu dugu, 3.2.1. ataleko parametro berberekin. Kasu honetan, baina, RG datu-multzoan lortutako emaitzei so, $I = 2 \cdot 10^8$ testuinguru sintetikora igo dugu eta $1,1 \cdot 10^9$ tokeneko WordNet corpusa sortu.

Kontuan izan, testu-corpusarekin ez bezala, grafoaren tamainak eta testuinguru sintetiko kopuruak corpus sintetikoaren informazio eraginkorraren kopurua mugatzen dutela. Hala ere, aurretiaz burututako esperimentuetan testuinguru kopurua handitu dugu, eta RG datu-multzoko emaitzetan hobekuntzarik lortu barik.

Datu-multzoak

Hainbat antzekotasun datu-multzo erabili ditu ebaluaziorako. Aurreneko hirurak antzekotasun hutsekoak dira: RG (Rubenstein and Goodenough, 1965), SimLex999 (SL) (Hill *et al.*, 2015) eta WordSim353 Similarity (WSS) (Agirre *et al.*, 2009a). Gainontzeko laurak ahaidetasun datu-multzoak dira: WordSim353 Relatedness (WSR) (Agirre *et al.*, 2009a), MTURK287 (MTU) (Radinsky *et al.*, 2011), MEN (Bruni *et al.*, 2014) eta WordSim353 osoa (WS) (Gabrilovich and Markovitch, 2007).

4.2.2 Emaitzak

Emaitzen atal honek emaitzen ebaluaziorako irizpideak laburbiltzen ditu lehenik. Ondoren, bi iturri konbinatzerakoan izandako hiru faseak deskribatzen ditugu; lehenengoan, testutik erauzitako hitz-bektoreen eta ezagutza-baseetako errepresentazio trinkoen emaitzak (SG_{wbu} eta AISG_{wn}, hurrenez hurren) 4.1. kapituluan azaldutako konbinaketekin konparatzen ditugu; bigarrenean, konbinaketa horietatik hoberena eta orduko artearen egoerako *retrofitting* fintze-metodoaren emaitzekin konparatzen dugu; hirugarrenean, gure metodoen emaitzak testu corpusetako eta ezagutza-baseetako informazioa uztartzen dituzten beste hiru metodorekin konparatzen ditugu.

Ebaluazioa

Ohiko praktikari jarraiki, datu-multzo bakoitzerako emaitzak Spearman korrelazioaren bidez ebaluatu ditugu. Gainera, antzekotasun hutseko datu-multzoen (RG, WSS, SL), ahaidetasun datu-multzoen (WSR, MTU, MEN) eta datu-multzo ororen (RG, SL, MTU, MEN, WS) emaitzen batezbestekoak ere kalkulatu ditugu (*antz*, *ahai* eta *dena*, hurrenez hurren). Izan ere, WSS eta WSR datu-multzoak WSen azpi-multzoak dira, eta, horrexegatik, hain zuzen, WSS antzekotasun datu-multzoen batezbestekoetan, WSR ahaidetasunekoenean eta WS denenenean darabilgu.

Ohikoa den bezala, metodoen korrelazio-emaitzen arteko desberdintasunaren esangura-maila Fisheren z-transformazioaren (Press *et al.*, 2002, 14.5.10 ekuazioa) bitartez neurtu dugu, %99ko konfiantza-mailarekin.

Testuan eta Wordneten oinarritutako hitzen errepresentazioen konbinaketa

Kapitulu honetako aurreneko esperimentuan, 4.3. taulan WordNeten (AISG_{wn}, lehen lerroa) eta testuan (SG_{wbu}, bigarren lerroa) oinarritutako hitz-bektoreen eta azken bi horien hainbat konbinaketen emaitzak azaltzen dira. Alde batetik, bi metodo horiek ahaidetasunean antzeko emaitzak erakusten dituzte, eta, beste aldetik, emaitzek WordNeteko informazioak antzekotasun-emaitzen hobekuntza eragiten dutela iradokitzen dute.

Are garrantzitsuago oraindik, 4.3. taulak SG_{wbu}rekiko irabaziak erakusten ditu 4.1. atalean azaldutako konbinaketa bakoitzerako. Bada, KAT eta ONA konbinaketak beste guztiei gailentzen zaizkie (*antz*, *ahai* eta *dena*), BB eta HIB dire hurrengo hoberenak, eta gainontzekoak okerrenak (RNK eta KNP

konbinaketak SG_{wbu} baino okerragoak). *dena* zutabeari dagokionez, ONA, KAT eta BB beste metodo guztiak baino esanguratsuki hobeak dira (%99ko konfiantza maila), baina azken hori ez da betetzen beraien artean. Datu hau kontuan hartuko dugu 4.3. ataleko esperimentuetan, BBren eraginkortasuna gure bi metodo hoberenen parekoa dela erakusten baitu, bi espazio baino gehiago konbinatzeko modu oso errazean ahalbidetuz.

Aipagarriak dira KKAren emaitzak, espazio elebakar bereiziekin lan egitean ez bezala (Faruqui and Dyer, 2014), testu eta WordNet informazio konbinatzerakoan hobekuntza ahula eragiten baitu. Izan ere, KATeko hobekuntzek azken bi horiek osagarriak direla erakusten dute, eta KKAn ikusitakoa ez dator bat ondorio horrekin.

ONA KATi gailentzen zaio, 600 dimentsioko bektoreak barik 300ekoak erabilita, 300 dimentsioko espazio konpaktua sortzea posible dela erakutsiz. Izan ere, 4.3. taulako emaitzei so, badirudi ONAk AISG_{wn} eta SG_{wbu} bektore-espazioetako informazioa modu eraginkorrean uztartzen duela (KATen antzera), bi espazio horiek gainditu egiten baititu. Gauzak horrela, ONArekin lortutako emaitzek testu eta WordNet bektore-espazioetako dimentsio batzuk korrelazioa dutela iradokitzen dute, esperimentuan erabilitako hitzen errepresentazio trinkoetan, behintzat.

Retrofitting metodoarekin konparaketa

Retrofitting metodoaren (Faruqui *et al.*, 2015) konparaketak arreta handiagoa merezi du, azken horrekin testu hutseko hitz-bektoreei WordNeten informazioa sar baitakioke, besteak beste. *Retrofitting* 2.3.2. atalean deskribatutako fintze-metodoen familiakoa izaki, bere sarrerak hurrengoak dira: alde batetik, aurretiaz ikasitako testu hutseko hitz-bektoreak; beste aldetik, grafoko hiztegiko hitz bakoitzarekin erlazionatutako hitzen (grafoko bere auzokide hurbilenak) zerrenda. Algoritmoak, aipatutako zerrendako hitzetatik abiatuta, jatorrizko testu hitz-bektoreak berrantolatzen ditu, erlazionatutako berbak hurbilduz. Faruqui *et al.*-en (2015) lanean, auzokideak lortzeko WordNeteko sinonimia eta hiperonimia erlazioak (WN_{sh}) ustiatzen dituzte, eta biak batera erabilita emaitza hoberenak lortzen. Gure WordNet metodoak sinonimia eta hiperonimia baino erlazio gehiagoz (glosa erlazioak barne) baliatzen gara, antzeko hitzen zerrenda gehigarria (WN_{all})[4] sortu dugu. Gauzak horrela, ze-

[4]Kontuan izan Faruqui *et al.*-ek (2015) bere artikuluan WordNeteko sinonimoz eta hiperonimoz osatutako zerrendari WN_{all} deitzen diotela, eta guk, ordea, WordNeteko erlazio guztiak erabiltzen ditugunean esleitzen diogula izen hori. Beraz, euren WN_{all}

	RG	SL	WSS	WSR	MTU	MEN	WS	antz	ahai	dena
$AISG_{wn}$	82,3	52,5	76,2	58,7	62,1	75,4	68,7	70,3	65,4	68,2
SG_{wbu}	76,4	39,7	76,6	61,5	64,6	74,6	67,3	64,2	66,9	64,5
KAT	7,8	**12,5**	6,7	6,5	7,5	**6,0**	8,0	9,0	**6,7**	8,4
ZEN	4,6	9,6	2,7	-1,1	1,3	3,2	2,3	5,6	1,2	4,2
KMP	-3,4	-1,2	-2,9	-8,9	-7,4	-0,9	-6,9	-2,5	-5,7	-4,0
ONA	**10,8**	**12,5**	5,7	5,3	**8,3**	5,6	6,9	**9,6**	6,5	**8,9**
KKA	6,8	2,7	-0,4	-0,2	11,7	-6,1	-3,5	6,0	-3,3	2,3
HIB	6,6	8,2	**7,2**	8,8	3,3	4,1	**8,6**	7,4	5,4	6,2
BB	8,0	12,1	5,5	6,5	7,0	6,2	7,4	8,5	6,6	8,2
RNK	7,3	11,3	0,2	**11,7**	-14,7	-14,7	6,6	6,2	-5,9	-0,8

4.3 taula – Testuko eta WordNeteko informazioaren konbinaketekin lortutako Spearman balioak. Lehengo lerroan WordNeten oinarritutako metodoaren emaitzak ($AISG_{wn}$), bigarrenean testuaren oinarritutakoarenak (SG_{wbu}), eta gainontzekoetan WordNeteko informazioa testuko informazioarekin konbinatzerakoan SG_{wbu}rekiko lortutako irabazi absolutuak. Ezkerrerengo zutabeek datu-multzo soilen emaitzekin, eta eskubirengo hiruak antzekotasun (*antz*), ahaidetasun (*ahai*) eta denen (*dena*) batezbestekoekin. Zutabe bakoitzeko emaitzarik hoberena letra lodiz.

rrenda horretako xede-hitz bakoitzari WordNeten zuzenean erlazionatutako (erlazio oro hartzen ditugu kontuan) hitz guztiak esleitzen dizkiogu.

4.4. taulako lehen hiru lerroetako emaitzek Faruqui *et al.*-en (2015)[5] testu hutseko hitz-bektoreen (CB_{far}) gainean WordNeteko informazioaren eragina azaltzen dute. Bada, *dena* zutabeari dagokionez, CB_{far}+WN_{all} eta CB_{far}+WN_{sh} CB_{far} soila baino hobeak dira maila esanguratsua (%99ko esanguramailarekin). Gauzak horrela, lerro horietako emaitzek erlazio oro (+WN_{all}) ustiatzea sinonimia eta hiponimia (+WN_{sh}) soilik erabiltzea baino zerbait

gure WN_{sh} da.

[5]Autoreek `https://code.google.com/archive/p/word2vec/` estekan eskuragarri dauden hitz-bektoreen (CB_{gnw} legez izendatutakoa atal honetan) azpi-multzo bat erabili dute.

	RG	SL	WSS	WSR	MTU	MEN	WS	antz	ahai	dena
CB_{far}	74,8	43,7	74,1	61,0	69,9	68,0	65,6	64,2	66,5	64,4
$+\text{WN}_{sh}$	5,0	7,4	4,0	-1,1	-0,6	2,6	1,9	5,5	0,3	3,3
$+\text{WN}_{all}$	4,9	2,5	2,6	4,3	2,4	5,7	3,7	3,3	4,1	3,9
SG_{wbu}	76,4	39,7	76,6	61,5	64,6	74,6	67,3	64,2	66,9	64,5
$+\text{WN}_{sh}$	4,6	-12,2	-4,8	-18,6	8,0	-4,9	-2,7	2,6	-4,3	1,3
$+\text{WN}_{all}$	6,3	0,9	2,3	0,2	2,4	0,9	0,9	3,7	0,3	2,1
ONA (gurea)	**10,8**	**12,5**	**5,7**	**5,3**	**8,3**	**5,6**	**6,9**	**9,6**	**6,5**	**8,9**

4.4 taula – Spearman balioak Faruqui *et al.*-en (2015) erabilitako hitz-bektoreekin (CB_{far}, goiko lerroa), eta *retrofitting* bidez WordNeteko informazioaren bi aldaera ($+\text{WN}_{sh}$ and $+\text{WN}_{all}$) konbinatzerakoan lortutako irabazi absolutuak. Erdialdean, testu-corpusak Skip-gramekin prozesatuta (SG_{wbu}) izandako emaitzak ere sartzen ditugu, aurretik aipatutako *retrofitting* irabaziekin batera. Beherengo lerroan, konparaketa errazte aldera, konbinazio hoberenarekin (ONA) lortutako emaitzak. Zutabe bakoitzeko emaitzarik hoberenak letra lodiz.

hobea dela erakusten dute, oro har (*dena* zutabea). Horrez gain, erlazio guztiekin ahaidetasunean nabarmen hobetzen dira emaitzak (*ahai* zutabea) eta zerbait okertzen antzekotasunean (*antz* zutabea). Horrek zentzua dauka, sinonimia eta hiperonimia antzekotasunari lotuta baitaude, eta glosa erlazioak ahaidetasunari (Agirre *et al.*, 2009a).

Taula bereko hurrengo hiru lerroetan SG_{wbu} hitz-bektoreei *retrofittingekin* erdietsitako irabaziak azaltzen dira. Aurreko kasuan ez bezala, $\text{SG}_{wbu}+\text{WN}_{all}$ eta $\text{SG}_{wbu}+\text{WN}_{sh}$ ez dira SG_{wbu} baino hobeak modu esanguratsuan, eta, ondorioz, *retrofittingak* WordNeteko informazioaren etekin gutxi atera diezaioke. Irabaziok xumeagoak dira CB_{far}en lortutakoekin alderatuta, testu hutseko hitz-bektoreak sortzeko erabilitako metodoen eta parametroen desberdintasunak direla-eta[6]. Esanguratsuak ez badira ere desberdintasunak, aurreko kasuan agertzen den fenomeno bera errepikatzen da, $+\text{WN}_{all}$ aplikatzeak ekarpen handiagoa egiten baitu. Esanak esan, *retrofittingen* ekarpena guk proposatutako lau konbinaketa hoberenetatik oso urruti geratzen da. Adibidez, 4.4. taulako azken lerroan gure konbinaketa hoberenaren, ONAren,

[6] *Retrofittingen* autoreek antzeko aldakortasunaren berri ematen dute.

emaitzak daude, eta datu-multzo orotan gailentzen zaio *retrofittingi*.

ONA konbinaketarekin erdietsitako emaitzak aurreko bi paragrafoetako kasuekin alderatuz gero, guztiak baino hobea da esanguratsuki (%99 esangura-mailarekin).

Gure konbinaketen emaitza altuagoak hurrengo arrazoigatik izan daitezke: alde batetik, WordNeteko errepresentazio trinkoak hobeto jasotzen dituzte hitzen eta adieren ñabardurak; beste aldetik, ausazko ibilbideek zuzeneko erlazioetatik haratago doazen erlazioak eta antzekotasunak jasotzeko gai dira. Izan ere, *retrofittingek* zuzeneko erlazioekin soilik egiten du lan.

Adibidez, WS urre-patroian *physics-proton* hitz-pareak antzekotasun altua dauka (10etik 8.12), 45. antzekotasun altuena du sailkapen-balioan 353 paretik, eta, 1. sailkapenean *tiger-tiger* da, 10eko antzekotasunarekin. Errepresentazio bakoitzeko antzekotasun-balioak konparatze aldera, datu-multzoko sailkapen-balioari egingo diogu erreferentzia. Gainera, ebaluazioa Spearman-ekin egiten dugu, sailkapen-balioan oinarrituta dagoena. Gauzak horrela, WordNeten *physics* eta *proton* hitzak zuzenean erlazionatuta ez badaude ere, AISG$_{wn}$ eta KAT bektoreetan *physics-proton* parea sailkapenean altu dago (41. eta 42. postuan, hurrenez hurren), eta SG$_{wbu}$n, ordea, nahiko behean (170. postua). *Retrofittingek* ezin du hitz-pare hau gehiagorik hurbildu, WordNeten hitz horiek modu ez zuzenean erlazionatuta baitaude, eta *retrofittingek* erlazio mota horiek ez baititu kontuan hartzen. Izan ere, SG$_{wbu}$k baino sailkapen-balio baxuagoa (193. postua) esleitzen dio *physics-proton* pareari, agian, beste erlazio batzuen eraginez urrundu egiten dituelako.

Alde negatiboari dagokionez, *retrofittingek* PPDBko (Ganitkevitch *et al.*, 2013) informazio sinplea modukoa ere gehi dezake[7], hobekuntza nabarmenak eragiten dituena (Faruqui *et al.*, 2015). Gure kasuan, konbinaketak burutze aldera, lehenik beharrezkoa da baliabide bakoitzerako hitzen errepresentazioak sorta osoa kalkulatzea. PPDB-rekin proba batzuk egin genituen, baina, ez genuen errepresentazio esanguratsurik lortu.

Beste metodo batzurekin konparaketa

Gure proposamenak *retrofitting* metodoarekin konparatu ondoren, atal honetan hurrengo hiru metodoekin dugu: CLEAR (Halawi *et al.*, 2012), Multiview LSA (Rastogi *et al.*, 2015) (MVLSA) eta ProjectNet (Tian *et al.*, 2015) (PNET). PNET, *retrofitting* bezala, fintze-metodoen familiakoa da,

[7]PPDBn hitz- eta esaldi-pareek parafrasi probabilitateak dituzte esleituta.

eta CLEAR eta MVLSA 2.3.1. atalean deskribatutako bateratze-metodoen baitan sartzen dira. Kontuan izan, azken horietako sarrerak testu-corpusak direla, eta hitz-testuinguru agerkidetzetako eta ezagutza-baseetako erlazioen informazioa bateratzen dutela beren ikasketa-prozesuetan.

CLEAReko autoreek testuan oinarritutako hitz-bektore propioak kalkulatzen dituzte oinarri-lerro legez, 100 dimentsiotakoak eta *Yahoo! Answer corpus*[8] gainean entrenatutakoak (TST_{yahoo}). WordNet murriztapenez baliatuta erdietsitako emaitzarik hoberenak aukeratu ditugu, hiperonimo-, meronimo- eta sinonimo-bikoteen bidez gauzatutako murriztapenak, hain zuzen. MVLSAko autoreak Polyglot Wikipedia dataseteko (Al-Rfou *et al.*, 2013) ingelesezko ataletik abiatzen da, eta LSAn oinarritutako euren ereduan entrenatu dute ($LSA_{polywiki}$), 300 dimentsiotan. Eredu horrekin testu-corpuseko informazioa, besteak beste, WordNetekoarekin aberasten dute. PNETek Freebaseko (Bollacker *et al.*, 2008) informazio estrukturalarekin Skip-gram ereduarekin Wikipedia gainean[9] ikasitako hitz-bektoreak (SG_{wiki9}) fintzen ditu.

Bada, 4.5. taulak aipatutako hiru metodoen Spearman emaitzak agertzen dira, aurreko ataleko zazpi datu-multzoetarako. Metodo bakoitzak bi lerro ditu; lehengoan, testu hutseko hitz-bektoreen emaitzak agertzen dira[10], eta, bigarrenean, azken horiek ezagutza-baseko informazioarekin aberastu ondoren erdietsitako irabazia[11].

CLEAR eta MVLSA emaitzak ez dira guztiz konparagarriak, teknika bakoitzak bere testu errepresentazio oinarri-lerroa baitauka, eta WordNeteko erlazio azpi-multzo desberdinak erabiltzen baitituzte. Hala ere, konbinaketen metodoek datu-multzo guztietan CLEAR eta MVLSA teknikek baino irabazi absolutu handiagoak lortzen dituzte, eta hori gure proposamena kontuan izatekoa denaren adierazlea da.

PNET ere ez da gure konbinaketekin guztiz konparagarria. Wikipediako erlazioak Freebaseekoekin estuki erlazioanatuta izaki, SG_{wbu} Wikipediako erlazioekin aberastuz gero PNETekin konparagarria izango litzateke. Bada, 4.3. atalean ikusiko dugunez, Wikipediako erlazioak gehi hiper-estekak erabili ditugu Wikipediako errepresentazio trinkoak sortzeko ($AISG_{wiki}$). PNET eta gure metodoa konparatze aldera, SG_{wbu} eta $AISG_{wiki}$ hitzen errepresentazioekin lortutako emaitzak BB bidez konbinatu ditugu, 4.6 puntuko irabazi absolutua (SG_{wbu}ren 64.5en gainetik) lortuaz. Bada, SG_{wbu} eta $AISG_{wiki}$ ar-

[8] `https://webscope.sandbox.yahoo.com/`

[9] `http://mattmahoney.net/dc/enwik9.zip`

[10] CLEAR metodoan TST_{yahoo}, MVLSAn $LSA_{polywiki}$ eta PNETen SG_{wiki9}.

[11] CLEAR metodoan +CLEAR, MVLSAn +MVLSA eta PNETen +PNET.

	RG	SL	WSS	WSR	MTU	MEN	WS
TST_{yahoo}	—	—	—	—	69,2	—	74,4
+CLEAR	—	—	—	—	-0,5	—	2,3
$LSA_{polywiki}$	71,2	34,5	76,8	60,1	59,1	71,4	68,0
+MVLSA	9,6	9,4	2,4	3,4	3,8	4,4	2,1
SG_{wiki9}	—	—	—	—	—	—	64,7
+PNET	—	—	—	—	—	—	3,7

4.5 taula – Testu hutseko hitz-bektoreak ezagutza-baseetako informazioarekin aberasterakoan erdietsitako Spearman balioak, CLEAR, MVLSA eta PNET metodoekin. Teknika bakoitzerako, abiapuntuko testu hutseko hitz-bektoreen eta teknikek eragindako irabazien emaitzak erakusten ditugu. Xehetasun gehiago goiko paragrafoan.

teko konbinaketarekin PNET teknikarekin baino irabazi absolutu altuagoa dago, WS datu-multzoan, bederen.

4.3 Esperimentuak, bi iturri baino gehiago konbinatzen

Bi errepresentazioren konbinaketekin erdietsitako hobekuntzak kontuan izanik, beste errepresentazio batzuetara hedatu ditugu konbinaketak. Atal honetan erabilitako baliabideak ere 4.2. taulan agertzen dira.

4.3.1 Baliabideak

Testuan oinarritutako hitz-bektoreei dagokienez, SG_{wbu} errepresentazioez gain, `word2vec` eredu-multzoaren egileek publikatutako hitz-bektoreak (ik. 3.2.1. atala) jaitsi ditugu (CB_{gnw}). Esanahi-bektoreok WBU baino corpus handiago batetik erauzi dira, baina parametroak ez dira antzekotasun datu-multzo batean optimizatu. Corpus desberdinean entrenatuak izaki, CB_{gnw} eta SG_{wbu} hitz-bektoreak osagarriak dira.

Ausazko ibilbideetan oinarritutako metodoari jarraiki, $AISG_{wn}$ errepresentazio trinkoez gain, Agirre *et al.*-ek (2015) publikoki eskuragarri jarritako Wikipedia grafotik (*wikiEN*) corpus sintetikoa erauzi dugu (*wiki*), eta, on-

doren, Skip-gram bidez bektoreak kalkulatu (AISG$_{wiki}$). Bada, birbideratze-, desanbiguazio- eta kategori-orriak kenduta, V Wikipedia grafoko orri ororekin osatuta dago. Wikipedia grafoak *a1* eta *a2* orrien artean n ertz izango ditu, baldin eta *a1* orritik *a2* orrira eta *a2* orritik *a1* orrira hiper-estekarik badago. Ezagutza-base hori 2.955.154 erpinez (orri) eta 16.338.664 ertzez (hiper-esteka) osatuta dago. Gure esperimentuetan $5 \cdot 10^8$ testuinguru sintetiko (3.1. ataleko I parametroa) sortu ditugu, $4,4 \cdot 10^9$ tokeneko corpusa eratuz. Testuinguru kopura kenduta, AISG$_{wiki}$ errepresentazio trinkoetarako ez dugu parametro optimizaziorik burutu; AISG$_{wn}$ errepresentazio trinkoetan erabilitako Skip-gramen parametro berekin egin dugu. Besterik gabe, RG datu-multzoko emaitzak konbergitu arte corpusa handitzen joan gara.

Azkenik, 3. kapituluko esperimentuen atalean legez, UKBrekin kalkulatutako Personalized PageRank bektoreak (PPB) ere erabili ditugu (Agirre and Soroa, 2009; Agirre *et al.*, 2014), bai WordNeterako (PPB$_{wn}$) eta bai Wikipediarako (PPB$_{wiki}$)[12]. Corpus sintetikoak erauzteko erabilitako WordNet eta Wikipedia grafoak orain arte aipatutako berak dira, eta moteltze-faktorea ere 0,85 balio lehenetsian utzi dugu. Gogoan hartu, AISG$_{wn}$ eta AISG$_{wiki}$ errepresentazio trinkoek ez bezala, PPB$_{wn}$ eta PPB$_{wiki}$ bektoreek euren grafoak beste dimentsio dituztela; hots, $117 \cdot 10^3$ eta $3 \cdot 10^6$ dimentsio, hurrenez hurren.

4.3.2 Emaitzak

Esperimentu-sail honetan sei errepresentazio osagarritatik abiatzen gara, eta gure xedea aipatutako konbinaketak errepresentazio horiek batzeko erabili daitezkeen aztertzea da. Bi iturri baino gehiago konbinatu nahi izanez gero, baina, hurrengo bi faktoreak hartu behar dira kontuan: bektoreen dimentsionaltasunen desoreka (300 versus milaka) eta konbinaketa posibleen kopurua. Azkartasuna eta eraginkortasunaren arteko oreka lortze aldera, gure lau konbinaketa hoberenetatik KAT eta BB soilik erabili ditugu. ONA denei gailentzen bazaie ere, ez dauka desberdintasun esanguratsurik KAT eta BB konbinaketekin (ik. 4.2.2. atala), eta azken bi horiek errazagoak eta azkarragoak dira. KATen erabilera, baina, dimentsionaltasun bereko bektoreetara soilik mugatu dugu, aurretiaz eginiko esperimentuetan, dimentsionaltasun desorekatuetako bektoreekin emaitzak oso txarrak lortu baititugu.

4.6. taulako goiko lerroek aipatutako sei errepresentazio bereizien perfor-

[12] 3. kapituluan bezala, hemen ere notazioa sinplifikatu gura izan dugu, eta, ondorioz, Wikipediako PPBei PPB$_{wiki}$ deitu diegu.

	RG	SL	WSS	WSR	MTU	MEN	WS	antz	ahai	dena
(a) SG_{wbu}	76,4	39,7	76,6	61,5	64,6	74,6	67,3	64,2	66,9	64,5
(b) CB_{gnw}	76,0	44,2	77,8	60,0	65,5	74,6	68,1	66,0	66,5	65,6
(c) $AISG_{wn}$	82,3	52,5	76,2	58,7	62,1	75,4	68,7	70,3	65,4	68,2
(d) PPB_{wn}	85,7	49,3	69,4	44,1	54,5	66,1	56,9	68,1	54,9	62,5
(e) $AISG_{wiki}$	79,6	32,3	67,5	48,2	43,9	60,9	59,3	59,8	51,0	55,2
(f) PPB_{wiki}	88,6	29,2	80,7	62,1	64,5	74,1	72,7	66,2	66,9	65,8
KAT(ac)	84,2	52,2	83,3	68,0	72,1	80,6	75,3	73,2	**73,6**	72,9
KAT(ace)	**91,2**	51,4	80,4	64,0	66,4	78,4	73,6	74,3	69,6	72,2
KAT(abce)	**91,2**	51,6	80,7	64,2	66,7	78,6	73,8	74,5	69,4	72,4
BB(ac)	84,4	51,7	82,1	68,0	71,6	80,8	74,7	72,8	73,5	72,7
BB(ace)	89,5	52,6	82,4	68,2	71,2	81,4	75,9	74,8	73,6	74,1
BB(abce)	89,0	52,1	83,5	68,2	73,4	81,7	76,5	74,9	74,4	74,5
BB(-f)	89,4	54,1	84,0	68,6	73,7	82,1	76,9	75,8	74,8	75,2
BB(-e)	86,4	53,8	83,8	**69,3**	**74,0**	81,8	76,3	74,6	75,0	74,4
BB(-d)	89,9	52,9	84,0	68,8	73,5	82,0	77,1	75,6	74,7	75,1
BB(-c)	89,6	51,4	83,9	66,8	70,8	80,6	76,2	75,0	72,7	73,3
BB(-b)	89,9	55,3	83,7	69,1	71,6	82,0	77,0	76,3	74,3	75,2
BB(-a)	90,4	**56,6**	83,2	62,7	71,8	81,6	77,1	**76,8**	72,0	75,5
BB(DENAK)	90,2	54,7	**84,3**	69,1	73,7	**82,8**	**77,4**	76,4	**75,1**	**75,7**
artearen egoera	86,0	55,2	80,0	*70,0*	*75,1*	80,0	*85,0*	73,7	75,0	*76,3*

4.6 taula – Taulan bi errepresentazio baino gehiagoren konbinaketen Spearman balioak azaltzen dira, zazpi datu-multzorako. Goiko sei lerroetan errepresentazio bereiziekin erdietsitako emaitzak, bakoitzari letra bat eskaintzen zaio. Erdiko hamahiru lerroetan errepresentazio desberdinen arteko KAT eta BB konbinaketa batzuen emaitzak, errepresentazio bakoitza letra batekin adierazten da (-x agertzerakoan, x errepresentazioa izan ezik beste guztiak), metodo guztienak barne (DENAK lerroa). Azkenengo lerroan arteren egoerako sistema desberdinen emaitzak. Ezkerrerengo zazpi zutabeetan datu-multzo bereizietako emaitzak, eta azkenengo hiruretan antzekotasun (*antz*), ahaidetasun (*ahai*) eta datu-multzo guztien (*dena*) emaitzak. Zutabe bakoitzerako, gure metodoen arteko emaitzarik hoberena letra lodiz, eta, artearen egoera letra etzanez.

mantzia azaltzen dute. PPB$_{wiki}$k ditu ahaidetasunean emaitzarik hoberenak, eta, AISG$_{wn}$ antzekotasunean eta orohar gailentzen da. Taulako hurrengo hiru lerroek KAT konbinaketa hoberenak erakusten dituzte. Emaitzei so, KAT oso eraginkorra da bi errepresentazio konbinatzerakoan, baina, bi iturritik gora kateatzerakoan, ordea, ez. Are gehiago, hirurekin eta laurekin birekin baino Spearman baxuagoak lortu ditugu.

Gainera, BBerekin konbinaketa gehiago burutu ditugu; KATekin erabilitakoez gain, bost (-x ablazio lerroetakoak) eta sei iturrikoak (*DENAK* lerroa) ere sartu ditugu. Bi iturrirekin KAT BB baino hobeto badabil ere, BBeren emaitzak gora egiten dute iturriak gehitu ala. Izan ere, Spearman altuena sei errepresentazioren emaitzen batezbestekoarekin erdietsi dugu. Ablazio lerroetako emaitzek metodo guztiek performantzia orokorrean ekarpena dutela erakusten dute, sei metodoko konbinaziotik edozein metodo kenduz gero emaitzak okertu egiten baitira (4.6. taulako ablazio eta *DENAK* lerroak).

Azkenik, azkenengo lerroak datu-multzo bakoitzean momentu hartako emaitzarik hoberenaren berri ematen du: RGen Hassan and Mihalcea (2011) emaitza, SLen Goikoetxea *et al.* (2015), WSSen Baroni *et al.* (2014), WSRen Baroni *et al.* (2014), MTUn Halawi *et al.* (2012), MENen Bruni *et al.* (2014), WSen Halawi *et al.* (2012). Guk proposatutako konbinazioek artearen egoera gainditzen dute RG, SL, WSS eta MEN datu-multzoetan. Kontuan hartu azken lerroko *antz*, *ahai* eta *dena* zutabeetako balioak sistema hoberenetako emaitzen batezbestekoak direla; hots, existitzen ez den sistema batekoak. Hala ere, gure konbinaketek antzekotasunean gainditu egiten dituzte, ahaidetasunean berdindu eta performantzia orokorrean ia berdindu. Gauzak horrela, modu bereizian ikasitako errepresentazioen konbinaketa sinpleek ikerketa-ildo emankorra izan daitezkeela erakusten dute.

4.4 Ondorioak

Kapitulu honetan, hurrengoa ondorioztatu dugu: izaera semantiko desberdinetako espazio bereiziak konbinatzea fintze- eta bateratze-metodoak baino eraginkorragoa dela. Izan ere, gure proposamena bi fintze-metodo eta bi bateratze-metodorekin alderatu dugu, emaitzek gure proposamena metodo horien hurbilpena baino efizienteagoa dela erakusten dute. Gure proposamenaren sendotasuna 3. kapituluan azaldutako ausazko ibilbideetan oinarritutako metodoan datza, azken horiek grafoaren egitura fintze- eta bateratze-metodoetako murriztapenek baino askoz hobeto jasotzen baitute. Konbinazio

sinpleen emaitza onak, bada, testuan eta WordNeten oinarritutako errepresentazio trinkoen osagarritasuna dute oinarri. Gainera, sei errepresentazio bereiziren batazbestekoen konbinaketak (BB) hainbat datu-multzotan artearen egoera gainditzen duela erakutsi dugu. Esperimentu hauetako software eta datu oro publiko egin ditugu, baita WordNeteko errepresentazio trinkoak eta kateatutako hitz guztien errepresentazio hibridoak ere[13].

Ezagutza-baseetako informazioa hitzen errepresentazioen ikasketan sartzeko metodoei dagokienez, gure emaitzek beste bide bat iradokitzen dute. WordNet bezalako ezagutza-base aberatsen errepresentazioak modu bereizian ikasteak ikerketa-ildo interesgarria dirudi, eta baita independenteki ikasitako espazioetan oinarritutako espazioak ustiatzen dituzten konbinazio metodoetan ikertzeak ere.

Hurrengo kapituluan testu eta ezagutza-baseen osagarritasunari eta konbinaketei buruz ikasitakoak bi eletara hedatuko ditugu.

[13] `http://ixa2.si.ehu.es/ukb/`

5

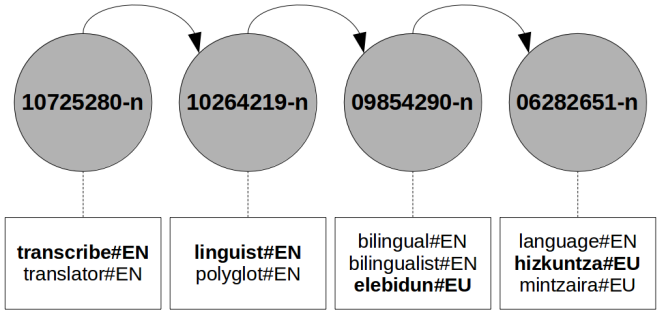
Eleaniztasunera hedapena

Aurreko bi kapituluetan testu eta ezagutza-baseetako informazio osagarria bateratzeko metodoak izan ditugu hizpide. 3. kapituluan deskribatutako ausazko ibilbideen metodoa ikerketa-ildo horren gakoa izan da, ezagutza-baseetako informazioa ustiatzeko bide eraginkorra dela erakutsi baitugu. 4. kapituluan testu eta ezagutza-baseetako espazio bereiziak konbinatzeak hitzen arteko antzekotasunean zein ahaidetasunean onura handiak dakartzala ikusi dugu.

Kapitulu honetako hipotesia hurrengoa da: espazio elebakarretan egindako aurkikuntza horiek guztiak espazio elebidunetan ere beteko dira. Hipotesi horretatik abiatuta, kapitulu honen xedea bektore-espazio elebidunak sortzea da, elearteko hitzen arteko antzekotasun- eta ahaidetasun-emaitzak hobetuko dituztenak. Ikerketaren fase honetara helduta, eta ausazko ibilbideen metodoak berebiziko garrantzia izan duela ikusita, eboluziorik naturalena azken hori elebidun bihurtzea izan da. Kapitulu honetan azalduko dugunez, ezagutza-base eleaniztunen azpiko egiturak[1] ustiatzeak ahalbidetuko ditu ausazko ibilbide elebidunak, eta azken hori elearteko antzekotasunaren eragile nagusi legez agertuko zaigu.

Wordnet eleaniztunen azpiko egitura ustiatzeko gure metodoa artearen egoerako mapaketa-metodoarekin konparatu dugu, sei hizkuntza-bikotetan eta hamabi elearteko antzekotasun eta ahaidetasun datu-multzotan, hurrengoak erakutsiz: 1) gure metodoak hiztegietan oinarritutako artearen egoerako

[1] Hizkuntzaren independentea, eta kontzeptu eta azken horien arteko erlazioez osatutakoa.



5.1 irudia – Wordnet 3.0 gaineko ingeles-euskara ausazko ibilbide elebiduna. Ibilbideak zeharkatutako *synsetak* grisez, eta azken horien ingelesezko (#EN marka) eta euskarazko (#EU marka) lexikalizazioak laukien barruan. Kontuan izan ezkerreko bi *synsetek* ez dutela lexikalizaziorik euskarazko wordneten. Ausaz aukeratutako lexikalizazioak urrats bakoitzean letra lodiz.

mapaketa-metodoa gainditzen du; 2) wordnet eleaniztunek[2], bere horretan, testuan espazioetan oinarritutako sistemak gainditzen dituzte antzekotasun datu-multzoetan; 3) testuko eta wordneteko informazioaren konbinaketa gakoa da.

5.1 WordNet eleaniztunak

Egun, wordnet publikoak hogeita hamalau hizkuntzatan daude[3]. Horietako batzuk jatorrizko ingelesezko egitura berbera jarraitzen dute (*expand* ikuspegia); hau da, kontzeptuek lexikalizazioak hainbat eletan dituzte, eta kontzeptuen arteko erlazioak bere horretan uzten dira. Beste wordnet batzuk, ordea, modu bereizian sortuak dira (*merge* ikuspegia), kontzeptu eta erlazio desberdinekin, baina, betiere ingelesezko WordNeterako mapaketa batekin. 2.2.1. atalean aipatu bezala, wordneten eleaniztasuna eta egitura oso baliotsuak dira hainbat atazatarako, eta tesi-lan honetan elearteko ahaidetasun- eta antzekotasun-emaitzak hobetzeko erabiliko dugu. Bada, kapitulu honetan proposatutako ezagutza-baseetan oinarritutako metodoa wordnet eleaniztu-

[2]Kapitulu honetan jatorrizko ingelesezko ezagutza-basea WordNet bezala izendatuko dugu, eta beste hizkuntza batzutan egindako egokitzapenak wornet bezala.

[3]`http://compling.hss.ntu.edu.sg/omw/`

netan oinarritzen da, eta, zehazki, wordnetak 3. kapituluan proposatutako metodoaren hedapena egiteko erabili ditugu.

5.1. irudiak WordNet 3.0g gaineko ausazko ibilbide elebiduna deskribatzen du, zehazki, ingelesa-euskara ele-bikotekoa. Bada, euskarazko wordneta (Pociello *et al.*, 2011) *expand* ikuspegiari jarraiki osatuta dagoenez, gure algoritmoak *synseten* gaineko ausazko ibilbideetan bi hizkuntzetako lexikalizazioak emititu ditzake. Hala, 5.1. irudian burututako ausazko ibilbide elebiduna hurrengoa da: *transcribe#EN linguist#EN elebidun#EU hizkuntza#EU*[4]. Corpus sintetiko elebidunean aipatutako lexikalizazioak soilik agertuko badira ere, berez, ausazko ibilbideak hurrengo *synseten* gainean egin da: *10725280-n 10264219-n 09854290-n 06282651-n*. Bada, adibidea horretan argi ikusten da ausazko ibilbidearen azpian WordNeteko informazio estrukturala dagoela, hizkuntzaren independentea dena, eta ele desberdinetako lexikalizazioen agerkidetzak WordNeten egitura horren emaitza direla. Gure proposamenaren muina, beraz, WordNeten egitura hori da.

5.2 Metodoa

Atal honetan bektore-espazio elebidunak sortzeko proposatutako bi metodo deskribatzen ditugu; ezagutza-baseetako erlazioak murriztapen elebidunen bidez ikasketa-prozesuan txertatzea eta ausazko ibilbide elebidunekin osatutako corpus sintetikoak. Bi metodoak independenteki erabili daitezke, baina, biak batera erabiliz gero, elearteko antzekotasun-emaitzak are gehiago hobetzen da.

Esanak esan, murriztapen elebidunen metodoa 2.4. ataleko bateratze metodoen multzoan sartzen da, Skip-gram ereduko (ik. 3.1.2. atala) galera-funtzioan ezagutza-baseko informazioa txertatuko baitugu. Corpus sintetiko elebidunak osatzeko algoritmoari dagokionez, 3.1.1. atalean deskribatutakoaren hedapena da.

5.2.1 Murriztapen elebidunak Skip-gram ereduan txertatzen

Lerrokatu gabeko bi corpus elebakarretik abiatuta, hitz-bektore elebidunak sortzeko modurik errazena corpusak esaldi mailan nahastea da, eta, ondoren,

[4]Ingelesezko lexikalizazioek #EN marka dute, eta euskarazkoek #EU.

corpus bateratu horretatik hitz-bektoreak kalkulatzea metodo distribuzionalen baten bidez. Metodo horren eraginkortasuna eleen artean partekatutako hitzetan datza, azken horiek ahalbidetzen baitituzte agerkidetza elebidunak (Wick *et al.*, 2016). Esaterako, zenbakiak (1,2 eta 3), izen propioak (*Hamlet* eta *Mandela*) eta mailegatutako hitzak (*soprano* eta *internet*) hizkuntzen artean berdinak direnez, corpus eleaniztun batean hitz horiek ele baten baino gehiagotan izango dituzte agerkidetzak. Zoritxarrez, partekatutako hitz horien kopurua murritzegia da, eta, ondorioz, itzulpenetako hitz baliokideak bektore-espazioan urruti egoteko joera dute. Aipatutako fenomeno hori enpirikoki erakutsiko dugu esperimentuen atalean.

Gauzak horrela, bektore-espazio bateratuan itzulpenen arteko antzekotasuna hobetzeko Skip-gram ereduko galera-funtzioari (ik. 52. orriko (3.3) ekuazioa) erregularizazio termino bat gehitzea proposatzen dugu, azken horrek itzulpeneko hitz baliokideak hurbil ditzan espazio bateratuan. Erregularizazio termino hori Halawi *et al.*-ek (2012) eta Bollegala *et al.*-ek (2016) erabili dute, eta, modu horretan, ohiko hitzen agerkidetzekin batera, kanpoko iturrietako informazio elebakar gehigarria txertatzen dute (sinonimoa, antonimoak, e.a.) ikasketa-prozesuan. Lan honetan azken metodo hori bi eletara hedatu dugu, hiztegi elebidunetatik ateratako itzulpenez baliatuta.

$$J_{sg+}(w,c) = J_{sg}(w,c) - \lambda \sum_{l=1}^{2} \sum_{m=1}^{M_l(w)} ||w - w_{lm}||_2^2 \tag{5.1}$$

(5.1) ekuazioak (w, c) hitz-testuinguru agerkidetza bakoitzerako galera-funtzio berria erakusten du; ezkerraldeko terminoa 52. orriko (3.3) ekuazioko J_{sg} agerkidetza bikoteen galera-funtzioa da; eskumakoa w hitzaren eta w_{lm} bere itzulpenaren arteko L2 norma minimizatzeko erregularizatzailea da, $\lambda \in \mathbb{R}^+$ erregularizazio terminoaren menpe dagoena. Hitz baten itzulpenaren baliokideak l (hizkuntza) eta m (l hizkuntzako $M_l(w)$ hainbat baliokideetako bat) bidez indexatuta daude. Kontuan izan, (5.1) ekuazioan l indizea 1etik 2ra doala, eta azken hori hitz baten murriztapenak bere hizkuntzan (murriztapen elebakarrak) eta beste hizkuntza batean (elearteko murriztapenak) izango direlako da. Esaterako, ingeles-euskara hizkuntza-bikotearekin lan eginez gero, WordNeten ingelesezko *witchery* hitzaren murriztapenak *witchcraft*, *aztikeria* eta *sorginkeria* dira, hots, aurrenekoa murriztapen elebakarra eta azkenengo biak eleartekoak. Gure esperimentuetan, murriztapen oro wordnetetatik atera ditugu (ik. 5.3.1. atala).

Kontuan izan, `word2vec` eredu-multzoak gradiente-jaitsiera estokastikoa

erabiltzen duela bere galera-funtzioa optimizatzeko (ik. 3.1.2), eta, ondorioz, atzeranzko propagazioan[5] gure murriztapenen gradienteak kalkulatu eta integratu behar direla. Atzerazko propagazioan 5.1 ekuazioko galera-funtzioaren gradientea kalkulatzen da c, w_n, w eta w_{lm} aldagaiekiko momenturo, eta dagozkien eguneraketak burutzen dira (xehetasun gehiago A.2 eranskinean). Atal honetan azkenengo biak deskribatuko ditugu, Skip-gram ereduko galera-funtzio originalarekin alderatuta, aurreneko bi deribatuak ez baitira aldatzen (ik. A.1 eranskineko (A.2) eta (A.4) ekuazioak).

Hala, logaritmoaren eta sigmoidearen deribatuak[6] kontua izanik, (5.2) ekuazioak galera-funtzio berriaren w aldagaiarekiko deribatua deskribatzen du:

$$\frac{\partial J_{sg+}}{\partial w} = (1 - \sigma(w^t c))c - 2\lambda \sum_{l=1}^{2} \sum_{m=1}^{M_l(w)} (w - w_{lm}) \tag{5.2}$$

(5.3) ekuazioak w_{lm} aldagaiarekiko deribatua azaltzen du,

$$\frac{\partial J_{sg+}}{\partial w_{lm}} = 2\lambda \sum_{l=1}^{2} \sum_{m=1}^{M_l(w)} (w - w_{lm}) \tag{5.3}$$

Hala, (5.2) eta (5.3) ekuazioetako gradienteak w eta w_{lm} aldagaiak eguneratzeko informazio gehigarria integratzen dute, jatorrizko Skip-gram ereduan existitzen ez zena. Kontuan izan, eguneraketa gehigarri horiek direla, hain zuzen, ezagutza-baseko informazio estrukturala txertatzen dutenak Skip-gram ereduko ikasketa-prozesuan. Gauzak horrela, Skip-gram ereduaren xedea hedatu egin dugu; hots, corpuseko hitz-testuinguru agerkidetzetako hitz-bektoreen arteko antzekotasuna indartzearekin batera, ezagutza-basean erlazionatutako hitzen arteko antzekotasuna ere sendotzen dugu.

Skip-gram ereduaren hedapen horrez gain, eta antzekotasun-emaitzak are gehiago hobetzeko asmoarekin, W eta C hitz- eta testuinguru-matrizeak konbinatu ditugu[7], beste lan batzuek azken horiek informazio osagarria dutela erakutsi baitute. Hala nola, `GloVe` (Pennington *et al.*, 2014) bektoreak osatze aldera, metodo horren autoreek W eta C batzea proposatzen dute. Azken

[5] W eta C matrizeak eguneratzeko urratsa.

[6] Bada, $\frac{\partial \sigma(x)}{\partial x} = (1 - \sigma(x))\sigma(x)$ eta $\frac{\partial log(x)}{\partial x} = \frac{1}{x}$, hurrenez hurren

[7] Gogoratu, 3. eta 4. ataletan C soilik erabili dugula.

horri jarraiki, Levy *et al.*-ek (2015) W eta C konbinatuta kalkulatutako kosinuak hitzen lehen eta bigarren graduko antzekotasunak kodetzen dituela ondorioztatu dute. Pennington *et al.*-ek (2014) eta Levy *et al.*-ek (2015) burututako ikerketek W eta C matrizeen batuketa azken horiek bereizita erabiltzea baino eraginkorragoa dela ondorioztatu dute. Horrexegatik, hain zuzen, eta aurreko bi kapituluetan ez bezala, kapitulu honetako esperimentuetan guk ere hitz-bektoreak W eta C matrizeetako bektoreen batuketa legez errepresentatuko ditugu.

5.2.2 Ausazko ibilbide elebidunak

3. kapituluan proposatutako metodoa eraginkorra da ezagutza-baseetako hitzen errepresentazioak erdiesteko, grafo baten gainean ausazko ibilbideen bidez lortutako agerkidetzekin ezagutza-basearen egitura kodetzeko gai baita. Bada, kapitulu honetan aipatutako metodoa bi eletara hedatu dugu.

Testuinguru sintetiko elebidunak sortzeko metodoa elebakarrean erabilitako berbera da (ik. 3.1. atala), baina, grafoko kontzeptu batetara heltzerakoan, bi hizkuntzetako lexikalizazioak emititu ahalko dira, ausaz. Metodo elebiduna 2. algoritmoaren xehetasunak 5.1. taulan daude formalizatuta. Gauzak horrela, K erpinak (kontzeptuak) dira, $N(k)$ k kontzeptuaren inguruko auzokideak dira grafoan (ertz baten bidez zuzenean lotutako erpinak), eta $D_l(k)$ $l \in \{l_A, l_B\}$ hizkuntzako k kontzeptuaren lexikalizazio guztien multzoa da. Kontuan hartu, l hizkuntza jakin bateko k kontzeptuak lexikalizaziorik ez badauka, $D_l(k)$ hutsik egon daitekeela. Izan ere, ele desberdinetako wordnetek estaldura desberdinak izateaz gain, kontzeptu berearen elearteko lexikalizazioetan desberdintasunak egon daitezke.

Kapitulu honetako esperimentuetan (ik. 5.3. atala) ingelesezko WordNeta, bi *expand* wordnet (euskara eta gaztelera) eta *merge* wordnet bat (italiera) erabili ditugu. 5.2 taulak wordnet bakoitzerako lexikalizazio eta *synset* kopuruak erakusten ditu. Taula horretan ikusten den bezala, wordnet horiek tamaina desberdinak dituzte, ingelesezkoa handiena izaki eta euskarazkoa txikiena.

Bada, *expand* ikuspegitik eratutako wordnetek kontzeptuen eta azken horien arteko erlazioen inbentarioa partekatzen dute, eta, hortaz, elearteko informazioa guztiz bateragarria da. Idealki, kontzeptu orok bi hizkuntzatan lexikalizatuta egon beharko luke, baina, errealitatean ez da hala. Izan ere, hizkuntza-bikoetako wordnet desberdinen estaldurak direla-eta, kontzeptuen frakzio bat baino ez dago lexikalizatuta bi eletan. *Merge* ikuspegian, elear-

Algorithm 2 Ausazko ibilbide elebidunak

Input: K kontzeptu multzoa
$N(k)$, $k \in K$ kontzeptuaren auzokideak grafoan
$D_l(k)$, $k \in K$ kontzeptuaren lexikalizazioak $l \in \{l_A, l_B\}$ elean
I, testuinguru sintetikoen kopurua
α, moteltze-faktorea
Output: SC, corpus sintetiko elebiduna
$CS \leftarrow []$
$i \leftarrow 0$
repeat
  $S \leftarrow []$
  $k \in K$ erpina aukeratu, $1/|K|$ probabilitatearekin
  repeat ▷ Ibilbidean jarraitu
    $l \in \{l_A, l_B\}$ hizkuntza aukeratu, $\frac{1}{2}$ probabilitatearekin
    if $|D_l(k)| > 0$ **then**
      $w \in D_l(k)$ hitza aukeratu, $1/|D_l(k)|$ probabilitatearekin, S
    end if
    $k' \in N(k)$ erpina aukeratu, $1/|N(k)|$ probabilitatearekin
    $k \leftarrow k'$
  until $random() > \alpha$ ▷ Ibilbidea geratu
  if $|S| > 0$ **then**
    $SC = SC \cup S$ ▷ Testuinguru berria
    $i \leftarrow i + 1$
  end if
until $i == I$

5.1 taula – 3. kapituluan proposatutako metodoan oinarritutako ausazko ibilbide elebidunen algoritmoa. Algoritmo horrek ezagutza-base eleaniztunetako lexikalizazioez baliatuta, elearteko agerkidetzez osatutako corpusa osatzen du.

	Lexikalizazioak	synsetak
wnEU	26.701	30.464
wnIT	46.679	49.515
wnES	53.039	55.814
wnEN	147.306	136.334

5.2 taula – WordNet bakoitzerako lexikalizazio eta synset kopurua, goranzko ordenean.

teko kontzeptuen eta erlazioen baliokidetasunak bilatze aldera, beharrezkoa da wordneten arteko mapaketa. Gure esperimentuetan italierarekin lan egiterakoan, *merge* ikuspegiaren inguruan esandakoak kontuan hartu ditugu. Gauzak horrela, italierazko eta ingelesezko wordneten kontzeptu oro grafo bateratu batean sartu ditugu, wordnet elebakar bakoitzeko erlazio guztiekin eta elearteko kontzeptuen mapaketarekin batera.

Hurrengo adibideak euskara (letra lodiz) eta ingelesa hizkuntzekin sortutako testuinguru sintetikoa erakusten du, *panthera* terminoa duena:

elur-pantera *panthera felidae sabertooth smiledon* ***ugaztun*** *erignathus*

5.3. taulak goiko adibide hori sortzeko 2. algoritmoak egindako urratsak deskribatzen ditu.

5.2.3 Corpus elebidunak sortzen

Hizkuntza-bikote bakoitzerako, izaera desberdineko hiru corpus mota erabiltzen ditugu:

- *tst*: testu-corpus elebiduna, bi corpus elebakarretako esaldien nahasketaz osatutakoa. Corpus hori bi eletako esaldi elebakarren bilduma da, eta ez dago esaldien arteko inongo asoziaziorik ez lerrokatzerik.

- *wn*: ezagutza-base batetik erauzitako corpus sintetiko elebiduna, 2. algoritmoa wordneten gainean aplikatuz lortua. Aurrekoak ez bezala, *wn* corpusak bi hizkuntzatik eratorritako hitzez osatutako esaldiak[8] ditu; hots, testuinguru sintetiko elebidunak.

[8]Termino hau ez dugu zentzu gramatikalean erabiltzen.

synsetak	lexikalizazioak	erlazioak
02128757-n	**elur-pantera#EU** ounce#EN panther_uncia#EN snow_leopard#EN	meronym member^{-1}
02128120-n	genus_panthera#EN **panthera#EN**	meronym member^{-1}
02120692-n	**felidae#EN** family_felidae#EN	gloss
02130545-n	saber-toothed_tiger#EN **sabertooth#EN**	gloss^{-1}
02130795-n	genus_smiledon#EN **smiledon#EN**	hyperonym
01864707-n	**ugaztun#EU** mamalio#EU mammal_genus#EN	hyponym
02080586-n	**erignathus#EN** genus_erignathus#EN	—

5.3 taula – Ausazko ibilbide elebidunen adibidea. Ezkerreko zutabean *synsetak*, erdikoan *synseten* lexikalizazio posible oro (euskarazkoak #EU bidez markatuak eta ingelesezkoak #EN bidez), eta eskubian ibilbideko hurrengo urratserako WordNet erpina. $^{-1}$ ikurrak alderantzizko erlazioa adierazten du, grafoa ez-zuzendua baita (ik. 2.2. atala). Synset bakoitzean ausaz aukeratutako lexikalizazioak letra lodiz.

- *hib*: corpus hibridoa aurreko bi puntuetan azaldutako *tst* eta *wn* corpusetako esaldiak nahastuz sortzen da. Corpus mota hau, hortaz, aurreko kapituluko 4.1.3. atalean azaldutako HIB konbinaketa da; hau da, corpus errealen eta sintetikoen konbinaketa, informazio distribuzionala eta ezagutza-baseetakoa bateratzen dituena[9]. Kapitulu honetan, HIB bi eletara hedatzeaz gain, token kopuruen kontrol zorrotzagoa aplikatuko diogu.

Bada, *tst* eta *wn* corpus elebidunetako informazio kopurua kontrolatze aldera, bikoteko hizkuntzek token kopuru bera dutela ziurtatu dugu. Era berean, aipatutako eleen arteko tokenen orekaz gain, *hib* corpusetan testu errealeko eta sintetikoko token kopuru berberekin osatu ditugu. Kontuan izan, 4. kapituluko HIB konbinaketan ez dugula azkenengo neurri hori kontuan hartu, *tst* eta *wn* corpusak inongo kontrol barik nahastu baititugu. Ikerketaren momentu hartan tokenen kontrolean sakondu ez bagenuen ere, esperimentuen atalean (ik. 5.3 atala) ikusiko dugunez, *hib* corpusak osatzeko *tst* eta *wn* corpusen token kopuruen oreka faktore garrantzitsua da.

Corpus mota bakoitzerako hitzen errepresentazioak kalkulatzeko, Skip-gram hedatua (ik. (5.1) ekuazioa) corpus horien gainean exekutatzen dugu.

[9] *hib* corpusetako hitzen errepresentazioei hitz-bektore hibrido deituko diegu.

Ereduak esaldi bakoitza isolamenduan prozesatzen du, xede-hitz bakoitzaren leihoak ez baititu inoiz esaldiaren mugak zeharkatzen. Esperimentuen atalean (ik. 5.3 atala) azalduko dugunez, ereduaren bi aldaerarekin egingo dugu lan; hots, ikasketa-prozesuan murriztapen elebidunak aplikatzen dituenarekin eta ez dituenarekin.

5.3 Esperimentuak

Aurreko bi kapituluetan legez, Spearman korrelazioaren (Spearman, 1904) bitartez ebaluatuko ditugu gure hitzen errepresentazio elebidunen kalitatea. Gainera, metodoen desberdintasunak (euren datu-multzo oro kontuan hartuta) estatistikoki esanguratsuak direla ikusteko, Demšar (2006) jarraitzen dugu; hau da, Wilcoxon test ez-parametrikoa erabiliko dugu datu-multzo bakoitzeko desberdintasunen gainean. Hala, metodo bikote bakoitzerako, aldebakarreko p-ren balio bakarra kalkulatzen dugu, eta, ohikoa den bezala, p-ren balioa 0,05 baliotik behera dagoenean desberdintasunak esanguratsutzat hartuko ditugu.

Esperimentuekin hasi aurretik, hainbat hizkuntzatako antzekotasun eta ahaidetasun datu-multzo aztertu ditugu, eta, ohikoa denez, ingelesak duela datu-multzo kopururik handiena. Gauzak horrela, xede-eleak aukeratze aldera, hurrengo faktoreak izan genituen kontuan: alde batetik, hizkuntzak bi datu-multzo (ik. 5.3.2. atala) izan behar ditu, gutxienez; beste aldetik, wordnet publikoa eduki behar du, eta kontuan hartzeko moduko tamainakoa izan behar da. Ingelesak, gaztelerak eta italierak betetzen dituzte bi faktore horiek, aurrenekoa mendebaldeko ele germanikoa, eta azkenengo biak erromantzeak. Aniztasun linguistikoa handitzeko, euskara aukeratu dugu, hizkuntza eranskaria eta isolatua. Egun, euskaraz kontuan hartzeko tamainadun wordneta badago, baina antzekotasun edo ahaidetasun datu-multzorik ez da existitzen. Gauzak horrela, guk geuk hurrengo datu multzoen euskarazko bertsioak eskuz sortu ditugu: WordSim353 eta RG. Hala, gure esperimentuetan lau hizkuntzarekin eta hamabi elearteko datu-multzorekin egin dugu lan. Horiek guztiak sei hizkuntza-bikotetan banatuta daude: ingelesa-euskara (ENEU), ingelesa-gaztelera (ENES), ingelesa-italiera (ENIT), gaztelera-euskara (ESEU), gaztelera-italiera (ESIT) eta euskara-italiera (EUIT).

Hitzen errepresentazioak, ebaluaziorako datu-multzoak (euskarazko elebakarrak barne) eta script oro publikoki eskuragarri daude, erreproduziga-

rritasuna errazteko argibideekin batera[10]. Hitzen errepresentazio elebidunak urratsez urrats hasieratik sortzeko jarraibideak ere sartu ditugu.

Hitzen errepresentazioak kalkulatzerakoan ez dugu Skip-gram parametrorik optimizatu, aurreko kapituluko parametroak berrerabili baititugu. λ erregularizazio koefizienteari dagokionez, Halawi *et al.*-en (2012) balio berarekin egin dugu lan. Laburbilduz, dimentsionaltasuna 300, 5 lagin negatibo, azpi-laginketa atalasea 0, leiho zabalera 5, eta λ 0,01. Aurreko bi kapituluetako esperimentuetan legez, ausazko ibilbideetako moteltze-faktorean berriro ere α=0,85 balio lehenetsia (Agirre *et al.*, 2010; Goikoetxea *et al.*, 2016) jarri dugu.

5.3.1 Baliabideak

Kapitulu honetako esperimentuetako metodoak eta baliabideak 5.4. taulan laburtu ditugu. Gure baliabide motak hurrengo laurak dira: elearteko datumultzoak (WS, SL eta RG), corpusak, murriztapenak (*mur*), eta wordnetak (*wnEN*, *wnEU*, *wnES* eta *wnIT*). Corpus motei dagokienez (ik. 5.2.3 atala), hiru aldaera daude. Aurrenekoa testu hutsa (*tst*) da, eta beste biak gure proposamenari dagozkio: WordNeten oinarritutakoa (*wn*) eta hibridoa (*hib*). Baliabideak elebakarrak zein elebidunak izan daitezke. Hala nola, *tstEU* eta *tstEUIT* euskarazko testu-corpus elebakarrarentzako eta euskara-italiera testu-corpus elebidunarentzako dira, hurrenez hurren.

Esperimentuetako hiru metodoak hurrengoak dira: gure oinarri-lerroa, bektore-espazio bereizien mapaketa lineala (Artetxe *et al.*, 2016) (MAP), Skip-gram corpus elebidunen gainean BAT eta azken horri WordNeteko murriztapenak aplikatzea (BATM). Hiru metodo horiek corpus mota bakoitzari aplikatzen zaizkie. Esaterako, BATM corpus hibridoaren gainean aplikatuz gero BATM_{hib} bezala izendatuko dugu.

Hitz-bektore elebakarrak ikaste aldera, Wikipedia corpusak[11] erabili ditugu. Euskara corpusa ezin izan dugu baliabide horretatik atera, ez baitago eskuragarri. Gauzak horrela, ingelesezko, gaztelerazko eta italierazko testu-corpusak dagozkien *dumpetatik* erauzi ditugu[12]. Euskarazko testu-corpusa 2016/04/07 Wikipedia *dumpetik* erauzi dugu, script propio bat erabilita, XML etiketak ezabatzen eta testu-edukia soilik gordetzen dituena. Corpus

[10] `http://ixa2.si.ehu.es/ukb/bilingual_embeddings.html`

[11] `http://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/`

[12] `http://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/` estekako script bat erabilita, eta taula eta matematika edukiak ezabatuz.

		Laburdura	Deskribapena
Baliabideak	Datu-multzoa	WS	WordSimS353 (ahaiadetasuna)
		SL	SimLex999 (antzekotasuna)
		RG	RG (antzekotasuna)
	Corpus mota	tst	Testu-corpusa
		wn	Wordnet corpus sintetikoa
		hib	Aurreko bien konbinaketa
	Ezagutza-baseak	wnEN	Ingelesezko Wordnet
		wnEU	Euskarazko wordnet
		wnES	Gaztelerazko wordnet
		wnIT	Italierazko wordnet
	Murriztapenak	mur	wordnetetako murriztapenak
Metodoak		MAP	Bektore-espazio bereizien mapaketa
		BAT	Skip-gram corpus elebidunetan
		BATM	Skip-gram murriztapenekin corpus elebidunetan

5.4 taula – Kapitulu honetako baliabideen laburdurak eta deskribapenak. Goiko hiru lerroek esperimentuotan erabilitako datu-multzoen laburdurak azaltzen dituzte (*WS*, *SL* and *RG*); hurrengo hiru lerroek corpus motak (*tst*, *wn* and *hib*), eta, ondorengo laurek wordnetak (*wnEN*, *wnEU*, *wnES* eta *wnIT*) eta azkenak murriztapenak (*mur*). Azkenengo hiru lerroetan hitz-bektore elebidunen metodoak agertzen dira; hots, mapaketa lineala MAP legez, Skip-gram corpus elebidunetan aplikatzean BAT moduan eta azken horri murriztapen elebidunak gehitzean BATM bezala. Azkenengo bi metodoak berriak dira, lan honetan proposatutakoak.

horren tamaina txikia dela-eta ($40 \cdot 10^6$ token), *Elhuyar Web Corpusarekin* (Leturia, 2012) osotu dugu, eskuragarri baitago autoreei eskatuz gero. Euskara hizkuntza eranskaria izaki, corpusari erro-bilatzaile propioa pasatu diogu (beste hiru eleetan ez). Corpus oro minuskuletara bihurtu ditugu. 5.5. taulak testu-corpus elebakarren tamainak azaltzen ditu.

Wordnetei dagokienez, *Multilingual Central Repository* 3.0 (Agirre *et al.*, 2012) baitako euskara, ingeles eta gaztelera wordnet irekiak erabili ditugu. 5.2.2. atalean esan bezala, euskarazko eta gaztelerazko wordnetek ingelesezko WordNet 3.0 bertsioaren *synset*, hierarkia eta erlazioa berrerabiltzen dituzte, glosa erlazioak barne. Italierarentzat ItalWordNet (Roventini *et al.*, 2003) erabili dugu, bere *synset* eta erlazio propiekin (Roventini *et al.*, 2003), ingelesezko WordNet 3.0 bertsiora estekak dituena. ItalWordNet eskaripean da, eta gainontzeko wordnetak *Multilingual Central Repository* (Agirre *et al.*, 2012) 3.0 errepositorioan daude eskuragarri[13]. Aurrerago esan bezala, 5.2. taulak

[13] `http://adimen.si.ehu.es/web/mcr/`

	Tokenak
tstEU	$160 \cdot 10^6$
tstIT	$380 \cdot 10^6$
tstES	$430 \cdot 10^6$
tstEN	$1.700 \cdot 10^6$

5.5 taula – Corpus elebakar bakoitzaren token kopurua, goranzko ordenean. Ingeleseko (tstEN), gaztelerako (tstES) eta italierako (tstIT) corpusak Wikipedia *dumpetatik* erauzita, eta Euskarazkoa (tstEU) bere Wikipedia *dumpaz* eta *Elhuyar Web Corpusaz* osatuta.

elebakar				elearteko		Totala
murEU	24.985	murIT	33.608	murEUIT	109.031	167.624
murEU	24.985	murES	43.218	murEUES	67.685	135.888
murES	43.218	murIT	33.608	murESIT	126.149	202.975
murEN	152.219	murEU	24.985	murENEU	99.861	277.065
murEN	152.219	murIT	33.608	murENIT	173.126	358.953
murEN	152.219	murES	43.218	murENES	176.873	372.310

5.6 taula – Hizkuntza-bikote bakoitzerako murriztapenak, goranzko ordenean. Hizkuntza-bikote bakoitzerako, murriztapen elebakarren, elearteekoen eta totalen berri ematen dugu.

wordnet guztien lexikalizazio eta *synset* kopuruak erakusten ditu. Corpusekin bateragarriak izate aldera, lexikalizazio oro minuskuletara bihurtu dugu. Kapitulu honetan jatorrizko ingelesezko ezagutza-baseari WordNet deituko diogu, eta, beste hizkuntzatakoei wordnet.

Murriztapenak sortzeko, wordnetetako hiztegiak ustiatu ditugu. Alde batetik, wordnet hiztegietako xede-termino bakoitzerako, azken horren sinonimo guztiekin osatutako murriztapen elebakarra sortzen dugu. Aipatutako sinonimoak xede-terminoaren *synset* guztietako lexikalizazioen baturatik ateratzen dira. Elearteko murriztapenak ere antzeko moduan sortzen ditugu; hots, ele jakin bateko xede-termino bakoitzerako, xede-terminoaren *synseten* beste hizkuntzako lexikalizazio guztien batura egiten dugu. Ele jakin bateko xede-termino baten murriztapen elebakarrak eta eleartekoak ditugunean, xede-termino horri murriztapenk denak esleitzen dizkiogu. 5.6. taulak hizkuntza-bikote bakoitzaren murriztapenak erakusten ditu.

Ebaluaziorako datu-multzoei dagokienez, elearteko datu-multzoak hurren-

go datu-multzo elebakarren ezagunen elearteko homologoak dira: WordSim353 (WS) ahaidetasun datu-multzoa (Gabrilovich and Markovitch, 2007), eta antzekotasun hutseko RG (Rubenstein and Goodenough, 1965) eta SimLex999 (SL) (Hill *et al.*, 2015). 2.3.1 atalean azaldu bezala, elearteko datu-multzoak Camacho-Collados *et al.*-ek (2015) proposatutakoa jarraitu dugu; hau da, datu-multzo jakin baten bertsio elebakarrak konbinatuta elearteko bertsioa sortu. Azkenengoa burutzeko erabilitako datu-multzo elebakarrak hurrengoak dira: Camacho-Collados *et al.*-ek (2015) sortutako italierazko WordSim353 eta SimLex999 eta gaztelerazko RG; Hassan and Mihalcea-ek (2009) eta Etcheverry and Wonsever-ek (2016) sortutako gaztelerazko WordSim353 eta SimLex999 datu-multzoak, hurrenez hurren; euskarazko WordSim353 eta RG, guk geuk sortutakoak[14]. Gauzak horrela, hizkuntza-bikote bakoitzarentzat bi elearteko datu-multzo eratu ahal izan ditugu. Ingelesa-gaztelera bikotean, baina, hiru sortzeko aukera izan dugu, eta euskara-italierakoan, ordea, bat bakarra. Denetara, hortaz, hamabi datu-multzo elearteko ditugu. WordSim353 datu-multzoaren kasuan 565etik 700era arteko bikote kopuruak ditugu (jatorrizkoak 352), SimLex999 datu-multzoan 1717tik 1982ra artekoak (jatorrizkoak 999) eta RGen 120tik 126ra artekoak (jatorrizkoak 65).

5.3.2 Emaitzak

Esperimentu guztietako emaitzak 5.7. taulan azaltzen ditugu. Emaitza horiek hiru azpi-ataletan aztertuko ditugu: hasteko, murriztapen elebidunen eragina testu hitz-bektoreen ikasketan aztertuko dugu (taulan ezkerrerengo *tst* hiru zutabeak); gero, ezagutza-baseetatik errepresentazio trinko elebidunak ikasteko gure metodoaren emaitzak (taulan erdialdeko hiru *wn* zutabeak); azkenik, corpus hibridoen bidez aurrekoak konbinatzeko gure metodoa (taulako hiru *hib* zutabeak). Hurrengo atalean emaitzok are gehiago aztertuko ditugu.

5.7. taulak NASARIren emaitzak ere erakusten ditu, elearteko antzekotasunean hainbat metodo gainditu dituena (Camacho-Collados *et al.*, 2016). NASARIren elearteko bektoreak erabili ditugu, publikoki eskuragarriak daudenak[15]. Ingelesezko, gaztelerazko eta italierazko bektoreak modu zuzenean eskuratu badaitezke ere, euskarazkorik ez dago eskuragarri. Autoreetako

[14] `http://ixa2.si.ehu.es/ukb/bilingual_embeddings.html`
[15] `http://lcl.uniroma1.it/nasari/`

			Gure lana								
		tst	tst		wn			hib			
		MAP	BAT	BATM	MAP	BAT	BATM	MAP	BAT	BATM	NASARI
WS	ENES	62,6	59,0	62,5	62,9	65,6	66,1	70,2	70,2	**72,0**	53,8
	ENIT	59,8	54,1	57,9	55,8	59,8	59,6	60,0	64,0	**64,9**	53,1
	ESIT	55,9	50,6	52,4	44,9	50,4	49,1	56,4	56,8	**58,8**	50,3
	ENEU	68,5	52,8	61,6	65,0	69,1	70,8	74,1	74,0	74,7	—
	ESEU	64,4	51,5	57,5	58,0	60,1	61,0	69,7	67,7	70,0	—
	EUIT	60,9	46,3	51,5	46,9	49,2	50,8	59,2	56,6	58,1	—
SL	ENES	33,9	20,5	26,9	46,1	47,4	48,2	46,5	48,6	**50,2**	44,0
	ENIT	35,8	25,0	32,0	42,8	46,9	46,9	40,4	46,3	47,9	**50,4**
	ESIT	31,6	25,1	28,7	38,5	42,7	42,7	40,3	40,7	43,3	**48,8**
RG	ENES	79,5	63,8	71,2	77,1	81,3	81,2	84,2	84,8	**85,7**	82,0
	ENEU	81,5	58,3	70,9	88,2	84,4	85,4	86,8	85,6	87,2	—
	ESEU	65,1	41,7	49,0	67,6	66,6	67,5	71,8	70,8	71,9	—

5.7 taula – Metodo eta datu-multzo guztientzako Spearman emaitzak. Ezkerrerengo zutabean MAP (Artetxe *et al.*, 2016) *tst* corpusen gainean aplikatzea lortutako emaitzak azaltzen dira; erdiko zutabeetan guk proposatutako corpusekin (*eb* and *hib*) edota metodoekin (BAT and BATM lortutakoak; eskubirengoan NASARI bektore bateratuekin (Camacho-Collados *et al.*, 2016) lortutakoak. Lerroetan, sei hizkuntza-bikoteetako elearteko datu-multzoak (WS, SL eta RG).

batek lagunduta, ingelesa-gaztelera emaitzak erreproduzitu ditugu, gainontzeko emaitzak ondo zeudela ziurtatuz. Kontuan izan, kosinuarekin barik, Camacho-Collados *et al.*-ek (2016) NASARI bektoreen antzekotasuna *weighted overlap* (WO) (Pilehvar *et al.*, 2013) metodoarekin kalkulatzen duela. Guk ere WO erabili dugu NASARI bektoreen elearteko antzekotasuna kalkulatzeko, eta gainontzeko metodoetan kosinua.

Lehenik eta behin, kontuan hartu 5.7. taulako emaitzak euren datu-multzo elebakar homologoetakoak baino zertxobait baxuagoak direla, halere, nahiko hurbil daude. Esaterako, Baroni *et al.*-ek, (2. taula) (2014) WSen 0,73 eta RGen 0,83 emaitzak lortzen dituzte CBOW testu hutsaren gainean aplikatzerakoan (atazen arteko konfigurazio hoberenarekin). Guk, ordea, 5.7. taulako *tst* zutabeetako metodoekin lortutako emaitzarik hoberenak 0,685 WSen eta 0,815 RGen dira, biak MAP_{tst} metodoarekin. Balio horiek ez dira konparagarriak, itzulpenek atazaren zailtasuna areagotu baitezakete eta balio horiek corpus eta parametro desberdinak erabilita lortu baitira (erreferentzia legez aipatzen ditugu soilik).

Hizkuntza-bikoteen artean konparaketa eginez gero, bikote batzuen emai-

tzak baxuagoak dira. Hala, ENEU bikoteak ditu hoberenak, eta ENIT eta ENES bikoteek antzeko balioak dituzte. Eleen arteko aldakortasun horretan hainbat faktorek eragiten dute, datu-multzoetako itzulpenetan polisemiaren eragina barne. Faktore horien azterketa ez da lan honen helburuetan sartzen. Bada, kapitulu honetan hainbat sistemen performantzien arteko konparaketak izango ditugu aztergai, betiere, sistema horien performantzia erlatiboa koherentea bada eleen eta datu-multzoen artean.

Murriztapen elebidunak testu-corpusen gainean

Aurreneko esperimentu-sailean testu hutsean zentratuko gara. Bada, corpus elebidunei wordneteko murriztapen elebidunak aplikatuta hitz-bektoreak kalkulatzeko metodoa (BATM_{tst}) eraginkorra den ala ez ikusiko dugu. Hizkuntza-bikote bakoitzerako, corpus elebiduna sortzeko Wikipedia *dumpetatik* erauzitako corpus elebakarren lerroak ausaz aukeratzen ditugu, gura dugun token kopurua lortu arte. Gero, esaldiak nahastuko ditugu, hizkuntza-bikote bakoitzerako corpus elebiduna osatuaz. Aipatutako corpus elebidunak Skip-gram ereduarekin prozesatu eta hitzen errepresentazio elebidunak kalkulatuko ditugu. Eleen artean corpus tamainak desorekatuta daudela-eta (ik. 5.5 taula), bi konbinaketarekin esperimentatu dugu: ele bietan token kopuru berarekin (corpus handiena duen elearen tamaina txikituta, lerroen ausazko aukeraketa bidez), edo ele bietako token guztiekin. Aurretiaz burututako esperimentuetan, lehenengo aukerarekin, token orekatuekin, emaitza konparagarriak edo hobeak izan ditugunez[16], metodo bateratuaren esperimentu guztiak corpus orekatuekin dira.

5.8. taulak testu hutseko corpus bateratuen tamainak erakusten ditu. Kontuan izan corpus bateratuetako testu elebiduna ez dagoela inondik inora ere lerrokatuta, eta esaldi bakoitza ele batetik datorrela. Oinarri-lerro legez, murriztapen gabeko hitz-bektore elebidunak ikasi ditugu (BATM_{tst}).

Gure hitzen errepresentazio elebidun oro artearen egoerako metodo batekin (Artetxe *et al.*, 2016) (MAP_{tst}) alderatuko dugu. Azken hori 5.5. taulako corpusetatik Skip-gram bidez ikasitako bektore-espazio elebakarretatik abiatzen da, eta wordnetetatik erauzitako hiztegi elebidunetan (ik. 2.4. atala) oinarritutako mapaketak ikasten ditu[17]. Gure metodoko hiztegi berberetaz

[16]Horrez gain, $5 \cdot 10^9$ tokeneko ingelesezko corpusarekin ere probatu dugu, baina hobekuntza barik.

[17]Kontuan izan metodo honek gureak baino token gehiagorekin egiten duela lan, emaitza hobeak lortu baititugu tamaina handiagoko corpusekin

	Tokenak
tstENEU	$320 \cdot 10^6$
tstESEU	$320 \cdot 10^6$
tstEUIT	$320 \cdot 10^6$
tstENIT	$760 \cdot 10^6$
tstESIT	$760 \cdot 10^6$
tstENES	$860 \cdot 10^6$

5.8 taula – Corpus elebidun bateratu bakoitzerako token kopurua, goranzko ordenean. Corpusak elebidunak esaldi elebakarren kopuru orekatuez (%50) osatuta.

baliatu gara (ik. 5.6. taula). Mapaketek norabide jakin bat daukate, wordnet handiena duen eletik txikiena duenera baitira, eta hizkuntza-bikote bakoitzerako bana egingo dugu, sei denetara. Hitzen arteko antzekotasuna kalkulatze aldera, xede-hizkuntzako hitzaren eta mapatutakoaren arteko kosinua kalkulatzen dugu. Esaterako, ingelesezko eta euskarazko hitz bikote baterako, mapatutako euskarazko hitzaren eta jatorrizko ingelesezkoaren arteko kosinua kalkulatzen dugu.

Testu-corpusetatik ikasterakoan, korrelazioak ausazko balioak baino gorago. 5.7. taularen ezkerrerengo hiru zutabeetan agertzen dira emaitzak, *tst* legez izendatuta. Oinarri-lerroak (BAT_{tst}), murriztapenik gabekoak, ausazkoak baino hobeak dira. Azken fenomeno hori ezustekoa da, eleen testuak independenteak baitira beren artean. Emaitzok Wick *et al.*-ek, (2. taula) (2016) erakutsitakoekin bat dator, antzeko konfigurazio batean 0,286 balioko kosinu-antzekotasuna lortu baitzuten. Autoreek fenomeno hori mailegu lexikalarekin lotzen dute; hots, eleen artean partekatutako zenbakiei, mailegatutako hitzei eta entitate izendunei (ik. 5.2.1. atala). Beraz, corpusen kateaketa soilarekin agerkidetza informazio nahiko batzen direnez, espazio bateratuak ikasterako orduan emaitza onargarriak erdiesten dira.

Testu-corpusetatik ikasterakoan, metodo elebidunen eragina. Murriztapen elebidunak (BATM_{tst}) txertatzeak BAT_{tst} hobetzen du, hizkuntza-bikote eta datu-multzo guztietan (p-ren balioa 0,0013). Azken horrekin guk proposatutako erregularizatzailea (ik. 5.2.1. atala) eraginkorra dela ondorioztatu dugu. Hala ere, Artetxe *et al.*-en (2016) mapaketa-metodoak (MAP_{tst}) BAT_{tst} eta BATM_{tst} metodoak gainditzen ditu datu-multzo guztietan (p-ren balioa 0,0013). Gauzak horrela, testu hutsarekin lan egiterakoan, egungo artearen egoerako metodoa (hitz-bektore elebakarrak modu bereizian

ikasi, eta, ondoren, mapaketa egin) gurea (corpusak elebakarrak kateatu, eta, ondoren, espazio bateratua ikasi hiztegi elebiduna murriztapen legez erabilita) baino hobea da. Hurrengo ataletan ikusiko dugunez, egoera irauli egingo da ezagutza-baseekin.

Ezagutza-baseetatik erauzitako corpus elebidunak

Esperimentu-sail honetan, wordneten gaineko ausazko ibilbideekin lortutako corpus sintetikoetan zentratuko gara (ik. 5.2.2. atala), testu hutseko corpusik erabili barik.

Aurreko atalean deskribatutako hiru metodoekin sortu ditugu hitzen errepresentazioak, baina, corpus sintetikoekin: artearen egoerako MAP_{wn}, eta guk proposatutako BATM_{wn} eta BAT_{wn}. Bada, MAP_{wn}eren kasuan, errepresentazio trinko elebakarrak corpus sintetiko elebakarretatik ikasi ditugu. Token kopurua konparagarria izate aldera (MAP_{tst}ekiko), 5.5. taulako tamainetako corpus sintetikoak sortu ditugu. Wordnet elebidunetik erauzitako hiztegiez baliatuta burutu ditugu mapaketak, aurreko atalean legez. BATM_{wn}en eta BAT_{wn}en kasuan, 5.8. taulako tamaina bereko corpus sintetiko elebidunak sortu ditugu. Atal honetan aipatu bezala, wordneten arteko tamaina desoreka dela-eta, ausazko ibilbide elebidunetan hizkuntza batek besteak baino token gehiago emitituko ditu. Hala nola, ENESen tokenen %65 ingelesez dira, eta %80ra arte igotzen da ENEU eta ENIT bikoteetan. Beste hizkuntzei dagokienez, ESEU eta ESIT bikoteetan %65 eta %70 token gazteleraz dira, hurrenez hurren, eta EUITen tokenen %60 euskaraz. BATM_{wn} metodoak MAP_{wn}eren murriztapen elebidun berdinak erabiltzen ditu. 5.7. taulan WN moduan izendatutako zutabeetan daude emaitzak.

Hiru metodoetako emaitzei so, BATM_{wn} eta BAT_{wn} metodoek MAP_{wn} Spearman balio altuagoa daukate hizkuntza-bikote eta datu-multzo guztietan (pren balioak 0,0067 eta 0,0034, hurrenez hurren), ENEUko RG datu-multzoan izan ezik. Bestalde, BATM_{wn} metodoa BAT_{wn}i gailentzen zaio hizkuntza-bikote eta datu-multzo gehienetan, baina desberdintasunak ez dira estatistikoki esanguratsuak. Beraz, corpus sintetikoen kasuan, espazio bateratu batetik ikasteak espazio bereizietatik ikasi eta azken horien mapaketa ikasteak baino emaitza hobeak ditu. Murriztapenekin hobekuntzarik erdietsi ez izanak zentzua dauka, corpus sintetikoek elebidunek inplizituki itzulpenak kodetzen baitituzte, eta, ondorioz, murriztapenak erredundanteak bihurtzen direlako, aurreko atalean ez bezala.

		tst	wn	Totala	tst	wn	Totala		
hibENEU	EN	32	128	320	266	114	760	ES	hibESIT
	EU	128	32		114	266		IT	
hibESEU	ES	53	107	320	72	308	760	EN	hibENIT
	EU	107	53		308	72		IT	
hibEUIT	EU	64	96	320	147	283	860	EN	hibENES
	IT	96	64		283	147		ES	

5.9 taula – Sei corpus elebidun hibridoetako token kopuruak (milioitan), ele pareen (lerroak) eta tokenen iturrien (zutabeak, *tst* testu naturala eta *eb* corpus sintetikoa) arabera banatuak.

Corpus hibrido elebidunak

Esperimentu-sail honetan, testu hutseko corpusak eta sintetikoak nahastuko ditugu (azken horiek corpus hibridoak dira, *hib* laburdura, ikusi 5.2.3. atala), eta, aurreko ataleko metodo berberak aplikatuko dizkiegu. BAT$_{hib}$ eta BATM$_{hib}$ metodoetan corpus hibrido elebidunak sortu ditugu, aurreko ataletako tamaina berekoak (ik. 5.8) eta eleen arteko tokenen orekak errespetatuz. Gauzak horrela, corpusen tamainak kontrolatu ditugu, eta, modu horretan, performantzia-desberdintasunak ez dizkiogu corpus handiagoak erabiltzeari egotziko. Atal honetan behin baino gehiagotan aipatu bezala, wordneten tamaina desberdinek corpus sintetikoetan eleen arteko tokeenen desorekak eragiten ditu. Aurretiaz burututako esperimentuetan, testu hutsetik eta ezagutza-baseetatik erauzitako token kopuru orekatuek emaitza hobeak lortzen dituztela ikusi dugu. Beraz, gure corpus hibrido elebidunek hurrengo baldintzak betetzen dituzte: 5.5. taulako token kopuru bera, eleen arteko token kopuru orekatua (%50 bakoitzerako), eta, baita testu hutseko eta testuinguru sintetikoko token kopurukoa ere. Baldintza horiei jarraiki, 5.9. taulak gure corpus hibrido elebidunen token distribuzioak erakusten ditu.

MAP$_{hib}$ metodoari dagokionez, bektore-espazio elebakarrak corpus hibrido elebakarretatik ikasi ditugu, *tst* eta *wn* corpusak konbinatzen dituztenak, hurrengo baldintzei jarraiki: alde batetik, corpus elebakar bereizien artean token kopuru orekatuak, eta, beste aldetik, corpusen elebakar bakoitzaren baitan token kopuru orekatuak (testu hutseko eta sintetikoko tokenen artean). Token kopuru osoak 5.8. taulakoak dira.

Proposatutako metodoen eragina corpus hibridoetan. 5.7. taulak

hiru metodoak corpus hibridoen gainean aplikatuta izandako emaitzak erakusten ditu (*hib* zutabeak), hizkuntza-bikote eta datu-multzo guztietarako. Kasu honetan, BATM_{hib} metodoak BAT_{hib} eta MAP_{hib} metodoek baino emaitza hobeak ditu hizkuntza pare eta datu-multzo orotan, EUIT pareko WS datu-multzoan ezik, MAP_{hib} baita hoberena. Bada, emaitza horiek murriztapenak konfigurazio honetan eraginkorrak direla iradokitzen dute, eta, orohar, gure BATM_{hib} metodoa MAP_{hib} (p-ren balioa 0,0043) eta BAT_{hib} baino hobea dela (p-ren balioa 0,0013). MAP_{hib} eta BAT_{hib} metodoen arteko desberdintasuna, baina, ez da estatistikoki esanguratsua. Laburbilduz, BATM_{hib} metodoak emaitzarik hoberenak ditu orohar, metodo eta hitzen errepresentazio guztien artean. Hala ere, datu-multzo bakun batzuetan gainditu egiten dute; hain zuzen, MAP_{tst} metodoan EUIT bikoteko WSen eta MAP_{wn}en ENEUko RGen. Zehazki, guk proposatutako BATM_{hib} metodoa (testu hutseko corpusak eta corpus sintetikoa konbinatzen duena) artearen egoerako MAP_{tst} baino eraginkorragoa da (p-ren balioa 0,0016).

5.4 Eztabaida

Atal honetako esperimentu kopurua eta lan honetan duten garrantzia kontuan izanik, ondorioen atalera igaro aurretik, izandako aurkikuntzen laburpena egingo dugu.

BAT espazioek ez dute MAP_{tst} oinarri-lerroa gainditzen. MAP_{tst} metodoaren BAT metodo kideek emaitzen analisi zehatzaren ondoren, hurrengoak ikusi ditugu: MAP_{tst} BAT_{tst} baino eraginkorragoa dela (p-ren balioa 0,0013), eta, testuinguru sintetikoak erabiltzerakoan kontrakoa gertatzen dena, BAT_{wn} metodoak MAP_{wn} gainditzen baitu (p-ren balioa 0,0067). BAT_{hib} eta MAP_{hib} metodoei dagokienez, ez dago desberdintasun esanguratsurik. Badirudi espazio bateratua testuinguru sintetikoekin eraginkorra dela (*wn*), eta mapaketa, ordea, testu hutsarekin (*tst*). Emaitza horiek espazio bateratua, bere horretan, ez dela mapaketa-metodoak gainditzeko nahikoa iradokitzen dute.

Ausazko ibilbideetan oinarritutako murriztapen elebidunek testu hutseko espazio bateratuak gainditzen dituzte; hala ere, ez dira MAP_{tst} gainditzeko gai. Aurreko paragrafoan esandakoaren antzera, BAT eta BATM metodoak alderatuz gero, BATM metodoak BAT metodoak baino Spearman korrelazio hobeak ditu, bai *tst* eta bai *hib* corpusetan (p-ren balioa 0,0013, bi kasuetan); hala ere, ez dago desberdintasun esanguratsurik

azken bi horien eta *wn* corpusen artean. Kontuan izan, BATM MAP baino hobea dela, bai *wn* eta bai *hib* corpusetarako (p-ren balioak 0,0034 eta 0,0043, hurrenez hurren). MAP_{tst}, baina, BATM_{tst} metodoari gailentzen zaio (p-ren balioa 0,0013) oraindik.

Ausazko ibilbide elebidunen erabilera. BAT_{wn} eta BATM_{wn} metodoek beren *tst* homologoak baino hobeak dira (p-ren balioa 0,0016 BAT metodoarentzat eta 0,0034 BATMentzat). MAP_{wn} eta MAP_{tst} metodoen artean, baina, ez dago desberdintasun esanguratsurik.

Ausazko ibilbide elebidunen ekarpena areagotu egiten da testu hutsarekin konbinatuz gero. Ausazko ibilbide elebidunen eragina areagotu egiten da testu hutsarekin konbinatzerakoan, *hib* corpusa darabilten hiru metodoek beren *tst* homologoak gainditzen baitituzte (p-ren balioa 0,0013 BAT eta BATM metodoetan, eta p-ren balioa 0,0027 MAPen). Gauzak horrela, testuinguru sintetikoak metodo bateratuarekin edo testu hutsarekin konbinatuz gero eraginkorra dela baieztatu dezakegu.

Ausazko ibilbide eta murriztapen elebidunak wordnet eleaniztunen gainean aplikatzea hiztegietan oinarritutako metodoak baino hobeak da. BATM_{hib} metodoak modu esanguratsuan gainditzen du MAP_{tst} datu-multzo guztietan (p-ren balioa 0,0016), EUIT hizkuntza-bikoteko WS datu-multzoan ezik. Beraz, alde batetik, azken horren bitartez wordnet elebidunak ez direla hiztegi soilak ondorioztatzen dugu, eta, bestetik, gure ausazko ibilbideen metodoak wordneten barne egitura ustiatzen dutela.

WordNeten oinarritutako corpus sintetiko elebidunek emaitzak hobetzen dituzte hitzen antzekotasun datu-multzoetan, baina ez ahaidetasunekoetan. Wordnet eleaniztunetako informazioa (testu hutsik erabili barik) testu hutsa eta hiztegi elebidunak baino eraginkorragoa da. Ausazko ibilbideetan soilik oinarritutako hiru metodoek (MAP_{wn}, BAT_{wn} eta BATM_{wn}) MAP_{tst} gainditzen dute sei antzekotasun datu-multzoetan (RG eta SL). Desberdintasuna estatistikoki esanguratsuak dira hiru kasuetan (p-ren balioa 0,0178 BAT_{wn} eta BATM_{wn} metodoetan, eta p-ren balioa 0,0296 MAP_{wn}en). Kontuan hartu sei datu-multzo soilik erabiltzen gabiltzala, eta Wilcoxon esangura-testa ahulagoa dela datu-multzo gutxirekin. Sei ahaidetasun datu-multzoetan (WS), ordea, ez dago desberdintasun esanguratsurik. Hala ere, BATM_{hib} metodoak MAP_{tst} gainditzen du (p-ren balioa 0,0296).

Murriztapen elebidunek eta WordNeten oinarritutako testuinguru sintetikoek artearen egoerako MAP_{tst} metodoa gainditzen dute hitzen ahaidetasun datu-multzoetan. Aurreko puntuan aipatu dugunez, BATM_{hib} metodoak MAP_{tst} gainditzen du, eta, ondorioz, wordnetak ustiatze-

ko gure metodoa wordnetak espazio bereizien mapaketetan erabiltzea baino eraginkorragoa dela ondorioztatu dugu. Horrez gain, 5.7. taulak BATM_{hib} metodoak NASARI (konparaketa posible den) 7tik 5 datu-multzotan gainditzen duela azaltzen du; hala ere, ez dago desberdintasun esanguratsurik beren artean. Kontuan izan zazpi datu-multzo soilik erabiltzen gabiltzala, eta Wilcoxon esangura-testa ahulagoa dela datu-multzo gutxirekin. Gainera, NASARIk BabelNet erabiltzen du, (hots, wordneten, Wikitionaryren eta Wikipediaren konbinaketa, besteak beste) eta guk wordnetak soilik. Gure aburuz, gure metodoa BabelNeteko aberastasunarekin konbinatuta hobekuntzak areagotzeko potentziala dauka.

BATM_{hib} metodoak artearen egoerako MAP_{tst} gainditzen du elearteko hitz antzekotasunean hizkuntza pare orotan, EUITen izan ezik. WordNeten estaldura aldatu egiten da (ik. 5.2. taula) (euskarazkoa (EU) da txikiena, italierazkoak (IT) jarraitzen dio), eta wordnet txikietan gure metodoak eragin ahulagoa izatea espero genuen. Wordnet elebidunaren kalitatea faktore garrantzitsua da performantzian: izan ere, gaztelerazko eta euskarazko wordnetak ingelesezkoarekin estuki lerrokatuta daude (grafoko erlazio berberak partekatzen dituzte, ikusi 5.2.2. atala), eta italierazkoaren lerrokatzea ahula da. Horrexegatik, hain zuzen, emaitzen hobekuntza txikiagoak espero genituen italierarentzat gure metodoarekin. BATM_{hib} eta MAP_{tst} metodoen emaitzak eleen artean alderatzerakoan, hobekuntzak daude hizkuntza-bikote eta datu-multzo guztietarako[18], EUIT parean ezik (WS datu-multzoan soilik). Kontuan izan euskara eta italiera sartzen dituzten pareekin ere hobekuntzak ditugula, eta azken biak batzerakoan (EUIT) soilik ez dagoela hobekuntzarik. Etorkizunera begira, lerrokatze ahuleko wordnetak errepresentatzeko beste metodoren batetik abiatuta, azken hori eraginkorragoa den ala ez aztertu gurako genuke. Horrez gain, gure proposamena esangura-testen analisiaren bidez sendotze aldera, hizkuntza pare bakoitzerako ebaluazioa datu-multzo gehiagora hedatu nahi dugu.

Gure corpus sintetikoaz osatutako corpus hibridoa, bere horretan, nahikoa da artearen egoerako MAP_{tst} gainditzeko. Hiru metodoak corpus hibridoei aplikatzerakoan (MAP_{hib}, BAT_{hib} eta BATM_{hib}) artearen egoerako MAP_{tst} gainditzen dute (p-ren balioak 0,0027, 0,0043 eta 0,0016, hurrenez hurren), eta, hobekuntzarik handienak metodo bateratuari murriz-

[18]Datu-multzo jakinetako desberdintasunak ez dira esanguratsuak z-test esangura testari jarraiki. Gure aburuz, azken hori datu-multzoen tamaina txikiak eragiten du (126 pare soilik, ikusi 5.2.2. atala).

tapen elebidunak gehituz gero (BATM_{hib}) erdiesten dira.

5.5 Ondorioak

Aurreko bi kapituluetan testu hutseko eta ezagutza-baseetako informazioa bektore espazio elebakarretan uztartzeko metodoak aztertu ditugu. Azken horietatik abiatuta, kapitulu honetan hitzen hitz-bektore hibrido elebidunak sortzeko metodo berria aurkeztu dugu, azken horiek espazio bateratuan sartzen dituena. Hori lortzeko, murriztapen elebidunak eta corpus sintetiko elebidunak (biak ere wordnetetatik eratorriak) erabili ditugu ikasketa-prozesuan. Wordnet elebidunak hiztegi elebidunak sortzeko baliatu badaitezke ere, guk metodo eraginkorragoa proposatzen dugu, ausazko ibilbideen bidez wordnetetako informazio estrukturala erauzten baitu. Bada, gure BAT_{hib} eta BATM_{hib} metodoek artearen egoerako metodoa (MAP_{tst}) (Artetxe *et al.*, 2016) esanguratsuki gainditzen dute. Gure proposamena wordnet elebidunen gaineko ausazko ibilbideetan dago oinarrituta, eta wordnetak zeharkatu ahala lexikalizazioak bi eletan emititzen ditu. 4.1.3. ataleko HIB konbinaketan inspiratuta, corpus sintetiko elebiduna testu hutseko corpus elebakarrekin konbinatu eta Skip-gram ereduaren bitartez prozesatzen dugu, emaitza legez hitz-bektore hibrido elebidunak jasoz. Are hobekuntza handiagoak lortzen dira Skip-gram ereduaren galera-funtzioan murriztapen elebidunak gehituz gero.

Beste ikuspuntu batetik, hitz-bektore elebidunak sortzerakoan, wordnet elebidunak ustiatzeko gure metodoa wordnetekin hiztegi elebidunak osatzea baino efektiboagoa da. Ingelesezko WordNetera lerrokatutako wordnetak gero eta gehiago direla kontuan hartuta, eta gure metodoa wordnetdun edozein hizkuntza-bikoteri aplikatu dakiokeela jakinik, guk proposatutakoak potentzial handia dauka. Besteak beste, DBpedia eta BabelNet moduko ezagutza-base eleaniztun handiek eleen estaldura eta emaitzen hobekuntzak aregotu ditzakete. Izan ere, gure metodo hoberenak (BATM_{hib}) wordnet soilez baliatuta NASARIk BabelNetekin erdietsitakoak baino emaitza hobeak (Camacho-Collados *et al.*, 2016) ditu; gure metodoa BabelNet moduko baliabide aberatsagoekin konbinatuz gero, beraz, emaitzak hobetu ahalko genituzke. Domeinu jakinetako ezagutza-baseak (esaterako, medikuntzakoak) lagungarriak izan daitezke alor espezifikoetan lan egiteko.

Gure ikerketa honek bektore-espazio eleaniztunak sortzeko hainbat aukera irekitzen ditu, ezagutza-base handiagoez baliatu baitaiteke, edota hitz

errepresentazio elebidunetatik eleaniztunetara aise hedatu. Gure metodoa beste ezagutza-base batzuei aplikatzeaz gain, etorkizunean corpus konbinaketen inguruko esplorazioan sakontzea planeatzen dugu. Ausazko ibilbideen bidez hitzen errepresentazio trinko elebidunak sortzeko aurreneko metodoa izaki, hobekuntzarako lekua dago oraindik. Hala nola, wordnet eleaniztunak ustiatu gurako genituzke eta bi eletik gorako hitz-bektore eleaniztunak elebidunak baino eraginkorragoak diren ala ez ikusi.

Hitzen errepresentazio elebidunak, ebaluaziorako datu-multzoak eta scriptak publikoki eskuragarri jarri ditugu[19], atal honetako emaitza oro erreproduzitzeko jarraibideekin batera. Hitzen errepresentazio elebidunak urratsez urrats sortzeko jarraibideak ere sartu ditugu.

[19] `http://ixa2.si.ehu.es/ukb/bilingual_embeddings.html`

6

Ondorioak eta etorkizuneko lanak

Lan honen xedea algoritmoek hitzen arteko antzekotasun-emaitzak hobetzea da, eta, horretarako, hitzen errepresentazioetara jo dugu. Izan ere, zenbat eta hitzen errepresentazio hobeak kalkulatu, orduan eta hobeto erreproduzituko dugu konputazionalki antzekotasuna.

Errepresentazio semantiko hobeak kalkulatzeko abiapuntua hurrengo hipotesia izan dugu: *testu-corpusetako agerkidetza informazioa eta ezagutza-baseetako informazio estrukturala osagarriak dira, eta, biak uztartuz gero, hitzen errepresentazio hobeak lortuko ditugu.* Bada, lan honetako metodo eta proposamen ororekin ikerketa-ildo hori jorratu dugu, betiere hitzen arteko antzekotasuna iparrorratz legez hartuta.

Errepresentazio distribuzionalak kalkulatzeko Skip-gram ereduaz baliatu gara, eta ezagutza-baseen errepresentazioak grafoetan oinarritutako ausazko ibilbideen bidez egin ditugu, bi metodo horiekin lortzen baitira antzekotasun-emaitzarik hoberenak. Baliabideei dagokienez, testu-corpusak eta WordNet ustiatu ditugu. Hala, ikerketako aurreneko urratsean (ik. 3. kapitulua) WordNeteko corpus sintetikoak erauzteko algoritmoa proposatu dugu, eta, corpus sintetikotik abiatuta, Wordeteko lexikalizazioen errepresentazio distribuzionalak kalkulatu ditugu. Sortutako baliabide horiek lan honen oinarriak ezarri dituzte, eta, gainera, antzekotasun-atazan artearen egoeran lortu dugu. Gero, bi baliabide horiek konbinatzeko hainbat metodo eta errepresentazio distribuzional hibrido proposatu ditugu (ik. 4. kapitulua), eta orduko artearen egoera gainditu dugu. Azkenengo urratsean errepresentazio distribuzional hibrido horiek eleaniztun bihurtu ditugu (ik. 5. kapitulua), eta elearteko antzekotasunean artearen egoera lortu dugu.

Azkenengo kapitulu honetan ekarpen nagusiak laburbilduko ditugu, eta, ondoren, aipatutako hiru kapituluen ondorioen sintesia burutuko dugu. Azkenik, etorkizuneko lan posibleak izango ditugu hizpide.

6.1 Ekarpenak

Testu-corpusak eta ezagutza-baseak errepresentazio semantikoak lortzeko baliabide nagusiak izan dira aspaldidanik, baina, tesi honen hastapenetan bi baliabide horien informazioa konbinatzeko proposamenak oso urriak ziren. Oro har, garai hartan 2.3.1. atalean deskribatutako bateratze metodo gutxi batzuk baino ez ziren proposatu, hala nola, Halawi *et al.*-en (2012) CLEAR, eta hirukoteetan oinarritutako (Wang *et al.*, 2014, en) eta (Xu *et al.*, 2014, ren) algoritmoak. Gainera, proposamen horiek guztiak espazio elebakarretan ziren, eta azkenaldion soilik agertu da espazio eleaniztunetara hedatutako proposamena, hots, Mrkšić *et al.*-en (2017) *attract-reppel* algoritmoa[1].

Bada, tesi-lan honen ekarpenak aurreko paragrafoan aipatutako ikerketa-ildo horretan baitan kokatzen dira, eta bi ekarpen mota bereiziko ditugu: alde batetik, artearen egoerari eginiko ekarpen zientifikoak, eta, beste aldetik, publikoki eskuragarri jarritako baliabideak. Hurrengo puntuetan gure laneko ekarpen zientifiko nagusiak deskribatuko ditugu:

Ausazko ibilbideen algoritmo batez baliatuz, WordNeten informazio estrukturala inplizituki corpus sintetikoetako agerkidetzetan kodetu dugu. *Ausazko ibilbideen algoritmoa* UKB programa-bildumaren (Agirre *et al.*, 2009b) gainean inplementatu dugu. Bada, azken horren PageRank Pertsonalizatua moldatu dugu, ausazko ibilbideetan zeharkatutako kontzeptuen lexikalizazioak fitxategi batean idatziz (ik. 3. kapitulua). Aipatutako algoritmo horren oinarriak WordNet grafoa eta horren hiztegi (elebakarra) dira. Gauzak horrela, WordNeteko erpinen gaineko milioika ausazko ibilbide corpus sintetiko batean gorde ditugu.

WordNetetik erauzitako corpus sintetiko batetik abiatuta, azken horren hitzen errepresentazio trinkoak kalkulatu ditugu. WordNet corpus sintetikoko hitz-testuinguru agerkidetzak Skip-gram ereduarekin prozesatu ditugu, eta *WordNet errepresentazio trinkoak* kalkulatu. Garai hartako ezagutza-baseetako hitzen errepresentazioekin alderatuta, gure

[1] Guk dakigunez, egun ez dago beste proposamenik.

hitz-bektoreek emaitza hobeak lortu dituzte, baina, dimentsionaltasuna eta konputazio-denbora nabarmen murriztuz.

Errepresentazio distribuzional hibridoak sortzeko metodo eraginkorrak proposatu ditugu. Lan honetan WordNeteko informazio estrukturala eta testuko hitz-testuinguru agerkidetzako informazioa konbinatzeko hainbat metodo proposatu ditugu (ik. 4. kapitulua) bektore-espazio elebakarretan. Azken horiek hurrengo multzoetan banatzen dira:

- *Bektoreen konbinaketa:* testu eta WordNet hitzen errepresentazioak kateaketaren, zentroidearen eta zenbaki konplexuen bidez konbinatu ditugu (ik. 4. kapitulua).
- *Korrelazio bidezko konbinaketa:* alde batetik, kateatutako testu eta WordNet hitzen errepresentazioei osagai nagusien analisia aplikatu diegu, eta, beste aldetik, azken bi bektore mota horiek korrelazio kanonikoen analisi bidez bektore-espazio partekatu batean ere proiektatu ditugu (ik. 4. kapitulua).
- *Corpusen konbinaketa:* testu-corpusak eta Wordnet corpus sintetikoak nahastu eta Skip-gram ereduarekin hitz-bektore hibridoak kalkulatu ditugu (ik. 4. kapitulua).
- *Emaitzen konbinaketa:* edozein hitz-errepresentazioren erdietsitako antzekotasun-emaitzen konbinaketak dira (ik. 3. eta 4. kapituluak); alde batetik, emaitzen batezbestekoak kalkulatu ditugu, eta, bestetik, emaitzen sailkapen-balioen batezbestekoak.

Aurreko guztien bertsio elebidunak landu ditugu. Azkenengo hiru ekarpenen bertsio hedatuak hurrengo puntuetan laburbiltzen ditugu, beste ekarpen gehigarri batekin batera:

- *WordNeteko corpus sintetiko elebidunak sortu:* WordNeteko egitura semantikoa hizkuntzaren independentea dela kontuan izanik, algoritmoa ausazko ibilbide elebidunak burutzeko egokitu dugu (ik. 5. kapitulua). Hala, ibilbideak WordNet elebidunekin burutuz gero, elearteko agerkidetzekin osatutako corpus sintetikoa lortzen dugu.
- *WordNeteko errepresentazio trinko elebidunak sortu:* Bektore-espazio elebidunak sortzeko hainbat metodo badaude ere (ik. 2.4. atala), gehiengo nagusia testu-corpusetan oinarritzen da. Gu, ordea, corpus

sintetiko elebidunetatik abiatu gara, eta, aurreko puntuan deskribatutako estrategiari jarraiki, errepresentazio trinko elebidunak kalkulatu ditugu (ik. 5. kapitulua). Guk dakigunez, hurbilketa hau planteatu duten bakarrak gara.

- *Errepresentazio distribuzional elebidunak sortu:* Corpus konbinaketak espazio elebidunetara hedatu ditugu (ik. 5. kapitulua). Bada, WordNeteko corpus sintetiko elebidunak testu-corpus elebakarrekin konbinatu eta Skip-gram bidez hitz-bektore hibrido elebidunak kalkulatu ditugu. Gainera, token kopuruen inguruko esplorazioa egin dugu; alde batetik, eleen arteko token kopuruen proportzioen eragina aztertu dugu; beste aldetik, testu-corpuseko eta WordNet corpus sintetikoko token kopuruena.

- *Murriztapen elebidunak txertatu ikasketa-prozesuan:* Skip-gram ereduari termino erregularizatzaile bat txertatu diogu (ik. 5. kapitulua). Hala, testuko hitz-testuinguru agerkidetzetako hitz-bektoreen antzekotasuna handitzeaz gain, behatutako hitzaren eta azken horrek WordNeten dituen *synseten* lexikalizazioen artekoa ere handitzen da. Modu horretan, grafoko informazio semantikoa ustiatzen dugu, hitz-bektoreen distribuzioa WordNeteko erlazioekin aberastuz. Corpus elebidunetatik abiatuta, WordNeteko *synsetetan* oinarritutako itzulpenen murriztapenak[2] txertatu ditugu Skip-gram ereduko ikasketa-prozesuan. Bada, hiztegietako itzulpenak erabili beharrean, WordNeteko egitura semantikoa ustiatu dugu, hizkuntzaren independentea baita.

Ekarpen zientifikoekin bukatu ondoren, publikoki eskuragarri jarritako baliabide guztiak zerrendatuko ditugu:

Ausazko ibilbideen metodoa: corpus sintetiko elebakarrak zein elebidunak sortzeko ausazko ibilbideen metodoa UKB 2.1 bertsioan dago inplementatuta[3].

Murriztapenekin hedatutako Skip-gram: jatorrizko `word2vec` eredumultzoko Skip-gramen galera-funtzioa murriztapenekin hedatu dugu eta `github`

[2]Kasu honetan ezin dugu sinonimia dela esan, berez, azken hori hizkuntza baten baitakoa delako.

[3]`http://ixa2.si.ehu.es/ukb/ukb_2.1.tgz`

errepositorioan dago eskuragarri[4].

Hitzen errepresentazio elebakarrak: lan honetako errepresentazio distribuzional elebakarretatik hurrengoak daude eskuragarri:

- *WordNeteko errepresentazio trinkoak:* glosadun WordNet 3.0 bertsiotik erauzitako errepresentazio trinkoak, formatu bitarrean ere eskuragarri[5] (ik. 3. kapitulua).
- *Kateatutako hitz-bektore hibridoak:* testu-corpus eta glosadun WordNet 3.0 bertsioko hitzen errepresentazioak kateatuta[6] (ik. 4. kapitulua).

Hitzen errepresentazio elebidunak: hurrengo baliabide oro 5. kapituluari dagozkio[7]:

- *Hitzen errepresentazioak:* bektore-espazio bateratuekin eta azken horiei murriztapenak gehituta kalkulatutakoak, testu eta ezagutza-base corpusekin eta bien konbinaketekin eta sei hizkuntza-bikoterako. Azken horiekin batera, mapaketa-metodoarekin (Artetxe *et al.*, 2016) kalkulatutako hitz-bektore elebidunak, corpus mota eta hizkuntza-bikote berdinentzat.
- *Hitzen errepresentazioak sortzeko urratsak:* deskribatutako hitzen errepresentazio mota guztiak hasieratik sortzeko urratsak.
- *Hitzen errepresentazioak sortzeko baliabideak:* erabilitako Hitzen errepresentazioak hasieratik sortzeko beharrezko baliabide guztiak, i.e. murriztapenak, mapaketarako hiztegiak, euskarazko erro-bilatzailea, wordnetak.
- *Publikazioko emaitza guztiak erreproduzitzeko scriptak.*

Urre-patroiak: hurrengo baliabide oro 5. kapituluari dagozkio[8]:

- *Elebakarrak*: euskarazko WordSim353 eta RG urre patroiak sortu ditugu, euskal hiztunen antzekotasun irizpideak erabilita.

[4] `https://github.com/JosuGoiko/word2vec_constraints`
[5] `http://ixa2.si.ehu.es/ukb/`
[6] `http://ixa2.si.ehu.es/ukb/`
[7] `http://ixa2.si.ehu.es/ukb/bilingual_embeddings.html`
[8] `http://ixa2.si.ehu.es/ukb/bilingual_embeddings.html`

- *Eleartekoak*: ingelesa, italiera, gaztelera eta euskara urre-patroi elebakarretatik abiatuta, Camacho-Collados *et al.*-ek (2015) proposatutako metodoari jarraiki, azken horiek beren artean konbinatu ditugu, eta, elearteko hamabi urre-patroi sortu.

6.2 Ondorioak

Gure errepresentazioen kalitatea hitzen antzekotasun edota ahaidetasun urre-patroien bidez ebaluatu dugu, eta emaitza horietatik ateratako ondorioek gure hipotesi nagusia zuzena dela ondorioztatu dugu, ikerketaren hastapenetik. Gainera, ondorio horiek elearteko bektore-espazioetan ere aplikatu daitezkeela ikusi dugu, eta, azken horrek are sendotasun eta orokortasun gehiago eman die ondorioei. Hala, lan osoan zehar ateratako ondorio nagusiak hiru puntutan laburbildu ditugu:

WordNet bektore-espazioak ezagutza-baseko informazio estrukturala kodetzeko gai dira. tesi-lan honen hastapenean WordNetetik erauzitako corpus sintetikoak osatzeko algoritmoa proposatu dugu (ik. 3. kapitulua). Skip-gram bezalako iragarpen-metodo batekin corpus sintetikoetako hitz-testuinguru agerkidetzak prozesatu daitezke, eta hitz-bektoreak kalkulatu. WordNetetik erauzitako corpus sintetiko batek gehienbat antzekotasun-erlazioak jasotzen dituela kontuan izanik, WordNeteko bektore-espazioak modu latentean erlazio horiek kodetzen ditu. Gauzak horrela, 3. eta 4. kapituluetan WordNeteko errepresentazio trinkoekin erdietsitako emaitzei so, azken horiek hitzen antzekotasun-atazan ahaidetasunekoan baino emaitza hobeak izateko joera erakusten dute. Are garrantzitsuagoa da, hala ere, aipatutako kapitulu horietako 3.5. eta 4.6. tauletako emaitzak islatzen dutena, WordNeteko errepresentazio trinkoak (AISG_{wn}) antzekotasun eta ahaidetasun emaitzek UKB bektoreenak (PPB_{wn}) berdindu edo hobetzen baitituzte.

5. kapituluan, elearteko ausazko ibilbideetatik kalkulatutako errepresentazio trinkoekin ere (BAT_{wn}) 5.7. taulan antzekotasun-emaitzak hobetzeko joera bera ikusi dugu. Emaitza horiek UKB bektore eleartekoekin alderatu ez baditugu ere, artearen egoerako NASARI bektoreen antzeko emaitzak lortu ditugu. Kontuan izan, NASARI bektoreak BabelNeten oinarritzen direla, eta azken hori Wikipediarekin eta WordNetekin osatu dagoela, besteak beste. Beraz, NASARI bektoreak eta gure WordNet errepresentazio trinkoak ez dira konparagarriak, baina, hala ere, pareko emaitzak lortu ditugu.

Zentzu horretan, gure proposamena askoz eraginkorragoa da, WordNet soilik ustiatuta antzeko emaitzak lortzen baititugu. Gauzak horrela, elebakarrak zein elebidunak izan, bektore-espazioek grafoen erlazioak kodetzeko gaitasuna dute.

Orotara, hurrengoak ondorioztatu ditugu: alde batetik, corpus sintetikoetako agerkidetzek WordNeten egitura semantikoa jasotzen dituztela[9]; beste aldetik, WordNet bektore-espazioak jatorrizko ezagutza-baseko egitura semantikoa modu eraginkorrean kodetzeko gai direla. Corpus sintetiko eta errepresentazio trinko elebidunekin ere ondorio berak atera ditugu.

WordNeteko eta testu-corpusetako informazio semantikoa desberdina baina osagarria da. 3. kapituluan bi baliabideak konbinatuz (ik. 3.5. taula) gure hipotesia egia dela ondorioztatu dugu aurrenekoz, testu eta WordNet errepresentazioen konbinaketak bi baliabideak bereizita baino emaitza hobeak baitituzte. Azken ebidentzia horri jarraiki, 4. atalean hainbat konbinaketa proposatu ditugu, eta emaitzak norabide berean doaz (ik. 4.3. eta 4.6. taulak). Bi iturri horien hitzen errepresentazioak edo corpusak konbinatuz gero hitzen errepresentazio semantikoa hobea da, eta, ondorioz, ahaidetasun- eta (batez ere) antzekotasun-emaitzak hobetzen ditu. Are gehiago, zenbat eta iturri gehiago konbinatu, badirudi esanahien are errepresentazio hobeak lortzen direla.

5. kapituluan elearteko murriztapenak eta corpus hibridoak (*hib*) deskribatu ditugu, eta aipatutako fenomenoa are argiago azaltzen da. Alde batetik, 5.7. taulan testu eta ezagutza-base corpus eta bien konbinaketaren (*tst*, *wn* eta *hib*, hurrenez hurren) emaitzetan azkena da garailea, metodo eta hizkuntza-bikote guztietan, salbuespenak salbuespen. Bestetik, ikasketa-prozesuan murriztapenak txertatzea onuragarria da testu-corpus eta corpus konbinatuetan[10] (*tst* eta *hib*, hurrenez hurren), testu eta WordNet informazioa konbinatzea ahalbidetzen baitute, antzekotasun-emaitzak hobetuz.

Orotara, hurrengoak ondorioztatu ditugu: iturri semantiko bereiziak konbinatzea azken horiek bereizita baino eta ahaidetasun- eta (batez ere) antzekotasun-emaitza hobeak lortzen dira, eta emaitzak are gehiago hobetzen dira iturriak gehitu ahala. Bi iturri baino gehiago erabili ez baditugu ere, bektore-espazio elebidunetan ondorio bera atera dugu.

Bektore-espazio hibridoak oso eraginkorrak dira antzekotasun-

[9]Beraz, corpus sintetikoak dira ausazko hitzez osatutako corpusak, inongo egitura ez informazio gabekoak.

[10]Gogoratu ezagutza-baseen corpusetan erredundantea dela, WordNet informazioa baitaukate soilik.

edota ahaidetasun-emaitzak hobetzerakoan. Azken hori ikerketa osoan presente izan dugu, hein handiagoan ala txikiagoan. Hala ere, garrantzitsua da gure metodoaren ezaugarri esanguratsuenetakoa nabarmentzea, ikerketa ildo bereko beste metodo batzuetatik bereizten baitu; bada, testutik erauzitako bektore-espazioa ezagutza-baseko informazioarekin aberastu barik (ik. 2.4. ataleko metodoak), bi espazioetako informazioa uztartzen dugu eta errepresentazio hibridoa sortzen. 4. atalean, esaterako, bi hurbilketa horiek alderatu ditugu, errepresentazio hibridoak[11] fintze- eta bateratze-metodoak baino eraginkorragoak direla ondorioztatuz. Izan ere, WordNeteko bektore-espazioek eta corpus sintetikoek fintze- eta bateratze-metodoetako murriztapen-zerrendek baino hobeto jasotzen dituzte ezagutza-baseetako erlazioen ñabardurak, eta, ondorioz, gailendu egiten dira ahaidetasun eta (batez ere) antzekotasun-atazan.

5. kapituluan errepresentazio hibrido elebidunak (BAT_{hib}) bateratze-metodoekin (elearteko murriztapenekin) bateratu ditugu (BATM_{hib}). Elearteko bektore-espazioetan, fintze- eta bateratze-metodoek barik, mapaketa-metodoek (MAP_{tst}) dute artearen egoera. Gauzak horrela, errepresentazio hibrido elebiduna bere horretan nahikoa da mapaketa-metodoa gainditzeko, eta emaitzak are gehiago hobetzen dira murriztapenak gehituz gero.

Orotara, hurrengoak ondorioztatu ditugu: testu eta ezagutza-baseetako espazioak uztartzea fintze- eta bateratze-metodoak baino hobea da antzekotasun-emaitzetan, aurrenekoek bi espazioen ñabardurak hobeto jasotzen baitituzte. Espazio elebidunekin ere ondorio berak atera ditugu.

Hala ere, bidean hainbat kale itsu aurkitu ditugu, eta frogatu beharrekoak ideiak batzuk ere bidean geratu dira. Hurrengoa atalean azken horiek izango ditugu hizpide, etorkizuneko lanak definituko baitituzte.

6.3 Etorkizuneko lanak

Tesi-lanetan lehentasunen, energiaren eta denboraren arteko oreka da galbaherik eraginkorrena, ikerketarako bidea garbitzen laguntzen baitu. Gauzak horrela, txosten honetan azken lau urteotako lan efektiboa deskribatu ditugu, hots, publikazioetan agertutakoa.

Egindako lanen artean badago hobekuntzarako eta hedapenerako tarterik. Gainera, ikerketa-ildo hau oso zabala da, eta hainbat alor esploratu barik

[11] Bai hitzen errepresentazioekin eta bai corpusekin osatutakoak.

utzi ditugu. Esanak esan, hurrengo puntuetan etorkizuneko lerro nagusiak deskribatuko ditugu:

Wikipedia gaineko ausazko ibilbideak hobetu. Wikipedia gaineko ausazko ibilbideek espero ez genuen portaera erakutsi dute, ezohiko lexikalizazioekin osatutako ibilbide anitz aurkitu baititugu. Gauzak horrela, 4. kapituluan Wikipediako errepresentazio trinkoekin ($AISG_{wiki}$) lortutako emaitzak Wikipediako UKB bektoreekin (PPB_{wiki}) erdietsitakoak baino okerragoak dira (ik. 4.6. taula). Fenomeno hori argitzeko hainbat esperimentu burutu baditugu ere, ez dugu ondorio esanguratsurik atera. Wikipedia grafoaren aberastasuna eta tamaina dela-eta, bere corpus sintetiko eta hitz-bektore egokiagoek ekarpen esanguratsuak egin ditzaketela uste dugu.

Ezagutza-basetik at dauden lexikalizazioak aberastu. Ikerketa ildo honetan arazo hau oso ezaguna da. Izan ere, fintze- eta bateratze-metodoek kontuan izan beharreko muga daukate, ezagutza-basearen baitako lexikalizazioen hitz-bektoreak soilik findu baititzakete. Gauzak horrela, testuko eta WordNeteko hitzen errepresentazioen mapaketekin saiakerak egin genituen, baina ez genuen emaitza esanguratsurik lortu. Etorkizunean muga hori gainditzea garrantzitsutzat deritzogu; alde batetik, Wordnet bezalako ezagutza-baseek estaldura nahiko murritza daukatelako, eta arazo hori konpontzeak emaitzak asko hobetuko lituzkeelako; beste aldetik, bi bektore-espazioen arteko erlazioa argituko genuelako. Kontuan izan, euskara bezalako hizkuntza gutxituek are muga gehiago dituztela zentzu horretan, eta, irtenbiderik aurkituz gero, etekin gehiago aterako diotela.

Bektore-espazio eleaniztunak sakonago aztertu. Tesi-lan honetan bektore-espazio elebidunak soilik aztertu ditugu, ez dugu gure ikerketa hizkuntza gehiagotara hedatu. Hala ere, hizkuntzen aniztasunaren eta kopuruaren eragina esploratzeak ekarpen interesgarriak egin diezazkieke bektore-espazio eleaniztunei. Aurreko puntuan aipatu bezala, baliabide urrieneko hizkuntzek aterako dute azken horretatik probetxu gehien, elearteko informazio-transferentziak beren gabeziak betetzeko gai direlako.

Euskarazko urre-patroi gehiago sortu. Lan honen xedea hitzen arteko antzekotasuna hobetzea izaki, esanguratsua iruditu zitzaigun euskarazko aurreneko urre-patroiak osatzea. Hala ere, urre-patroiak hasieratik sortzeak denbora dezente eskatzen du, eta lan honetan zehar bi egingarrienak osa-

tzeko erabakia hartu dugu: WordSim353 eta RG[12]. Urre-patroiek euskarazko hitzen errepresentazioen kalitatea neurtzeko baliabide esanguratsua dela kontuan izanik, etorkizunari begira halako gehiago osatzeko asmoa daukagu. Hautagaien artean (zinezko) antzekotasun SimLex999 urre-patroia da egokiena, erabilienetakoa izateaz gain, adjektiboen, aditzen eta izenen azpimultzoz osatuta dagoelako.

Ezagutza-baseetako egitura semantikoak espazio hiperbolikoetan kodetu. Lan honetan ezagutza-baseetako informazio semantiko oro espazio euklidear lauan kodetu dugu. Espazio hiperbolikoen geometria, baina, ezagutza-baseen hierarkiak kodetzeko egokiagoa da, eta azkenaldion hizkuntzaren prozesamenduan ildo berri hori jorratzen hasi direnak ere agertu dira (Chamberlain *et al.*, 2017; Nickel and Kiela, 2017). Espazio hiperbolikoek potentziala kontuan izanik, ezagutza-baseak azken horietan kodetzea etorkizun oparoko ikerketa-lerroa izan daiteke.

Ebaluazio estrinsekoak. Esperimentu guztietan, bektoreak antzekotasun-atazan ebaluatu ditugu, hots, ebaluazio intrintseko batean. Hala ere, azkenaldion hizkuntzaren prozesamenduan hitz-bektoreen erabilerak izandako gorakada dela-eta, ebaluazio estrinsekoek gero eta pisu gehiago hartzen dabiltza, hala nola, sentimendu-analisiak, dialogo-sistemak, rol semantikoen etiketatzeak edota entitate izendunen ezagutzeak. Gainera, egun ebaluazio intrinsekoen eta estrinsekoen arteko lotura ez dago argi, eta lotura hori esploratzeak ere ekarpen esanguratsuak egin ditzake hitz-bektoreen ebaluazioan.

[12]Kontuan izan, jatorrian elebakarrak badira ere, guk elearteko espazioetan soilik erabili ditugula.

Bibliografia

Agirre A.G., Laparra E., Rigau G., and Donostia B.C. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. *GWC 2012 6th International Global Wordnet Conference*, page 118, 2012.

Agirre E., Alfonseca E., Hall K., Kravalova J., Paşca M., and Soroa A. A study on similarity and relatedness using distributional and WordNet-based approaches. *Proceedings of of HLT-NAACL*, 19–27, 2009a.

Agirre E., Barrena A., and Soroa A. Studying the wikipedia hyperlink graph for relatedness and disambiguation. *arXiv preprint arXiv:1503.01655*, 2015.

Agirre E., Cuadros M., Rigau G., and Soroa A. Exploring knowledge bases for similarity. *LREC*, 2010.

Agirre E., De Lacalle O.L., Soroa A., and Fakultatea I. Knowledge-based wsd and specific domains: Performing better than generic supervised wsd. *IJCAI*, 1501–1506, 2009b.

Agirre E., López de Lacalle O., and Soroa A. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84, 2014.

Agirre E. and Soroa A. Personalizing pagerank for word sense disambiguation. *Proceedings of the 12th Conference of the European Chapter of the*

Association for Computational Linguistics, 33–41. Association for Computational Linguistics, 2009.

Al-Rfou R., Perozzi B., and Skiena S. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*, 2013.

Artetxe M., Labaka G., and Agirre E. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. *Proceedings of EMNLP*, 2289–2294, Austin, Texas, 2016.

Artetxe M., Labaka G., and Agirre E. Learning bilingual word embeddings with (almost) no bilingual data. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1 lib., 451–462, 2017.

Artetxe M., Labaka G., and Agirre E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, 2018.

Avrachenkov K., Litvak N., Nemirovsky D., and Osipova N. Monte carlo methods in pagerank computation: When one iteration is sufficient. *SIAM J. Numer. Anal.*, 45(2):890–904, 2007.

Banerjee S. and Pedersen T. An adapted lesk algorithm for word sense disambiguation using wordnet. *International Conference on Intelligent Text Processing and Computational Linguistics*, 136–145. Springer, 2002.

Banko M., Cafarella M.J., Soderland S., Broadhead M., and Etzioni O. Open information extraction from the web. *IJCAI*, 7 lib., 2670–2676, 2007.

Baroni M., Dinu G., and Kruszewski G. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of ACL (1)*, 238–247, 2014.

Barrena A., Soroa A., and Agirre E. Alleviating poor context with background knowledge for named entity disambiguation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1 lib., 1903–1912, 2016.

Bengio Y., Ducharme R., Vincent P., and Jauvin C. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Bian J., Gao B., and Liu T.Y. Knowledge-powered deep learning for word embedding. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 132–148. Springer, 2014.

Blei D.M., Ng A.Y., and Jordan M.I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Boeree C. Psychology: the beginnings. *Retrieved April*, 26:2008, 2000.

Bojanowski P., Grave E., Joulin A., and Mikolov T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Bollacker K., Evans C., Paritosh P., Sturge T., and Taylor J. Freebase: a collaboratively created graph database for structuring human knowledge. *Proceedings of the ACM SIGMOD*, 1247–1250, 2008.

Bollegala D., Alsuhaibani M., Maehara T., and Kawarabayashi K.i. Joint word representation learning using a corpus and a semantic lexicon. *Proceedings of AAAI*, 2690–2696, 2016.

Bowman S.R., Angeli G., Potts C., and Manning C.D. A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 632–642. 2015.

Brin S. and Page L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

Bruni E., Tran N.K., and Baroni M. Multimodal distributional semantics. *JAIR*, 49:1–47, 2014.

Budanitsky A. and Hirst G. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

Burnham W.H. Memory, historically and experimentally considered. i. an historical sketch of the older conceptions of memory. *The American Journal of Psychology*, 2(1):39–90, 1888.

Camacho-Collados J., Pilehvar M.T., and Navigli R. A framework for the construction of monolingual and cross-lingual word similarity datasets. *Proceedings of ACL (2)*, 1–7, 2015.

Camacho-Collados J., Pilehvar M.T., and Navigli R. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64, 2016.

Chamberlain B.P., Clough J., and Deisenroth M.P. Neural embeddings of graphs in hyperbolic space. *arXiv preprint arXiv:1705.10359*, 2017.

Chisholm A. and Hachey B. Entity disambiguation with web links. *Transactions of the Association of Computational Linguistics*, 3(1):145–156, 2015.

Chiu B., Pyysalo S., Vulić I., and Korhonen A. Bio-simverb and bio-simlex: wide-coverage evaluation sets of word similarity in biomedicine. *BMC bioinformatics*, 19(1):33, 2018.

Collobert R. and Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th international conference on Machine learning*, 160–167. ACM, 2008.

Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., and Kuksa P. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

Conneau A., Lample G., Ranzato M., Denoyer L., and Jégou H. Word translation without parallel data. *CoRR*, abs/1710.04087, 2017.

Cop U., Dirix N., Drieghe D., and Duyck W. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615, 2017.

de Melo G. and Weikum G. Menta: Inducing multilingual taxonomies from wikipedia. *Proceedings of the 19th ACM international conference on Information and knowledge management*, 1099–1108. ACM, 2010.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., and Harshman R. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

Demšar J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, 2006.

Etcheverry M. and Wonsever D. Spanish word vectors from wikipedia. In Chair) N.C.C., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., and Piperidis S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 3681–3685, Paris, France, 2016.

Faloutsos M., Faloutsos P., and Faloutsos C. On power-law relationships of the internet topology. *ACM SIGCOMM computer communication review*, 29 lib., 251–262. ACM, 1999.

Faruqui M., Dodge J., Jauhar S.K., Dyer C., Hovy E., and Smith N.A. Retrofitting word vectors to semantic lexicons. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1606–1615, Denver, Colorado, 2015.

Faruqui M. and Dyer C. Improving vector space word representations using multilingual correlation. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 462–471, Gothenburg, Sweden, 2014.

Finkelstein L., Gabrilovich E., Matias Y., Rivlin E., Solan Z., Wolfman G., and Ruppin E. Placing search in context: The concept revisited. *Proceedings of the 10th international conference on World Wide Web*, 406–414. ACM, 2001.

Firth J. A synopsis of linguistic theory 1930-1955 in studies in linguistic analysis, philological society, 1957.

Gabrilovich E. and Markovitch S. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. *Proceedings of IJCAI*, 7 lib., 1606–1611, 2007.

Galton F. I. co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*, 45(273-279):135–145, 1889.

Ganitkevitch J., Van Durme B., and Callison-Burch C. Ppdb: The paraphrase database. *Proceedings of HLT-NAACL*, 758–764, 2013.

Goikoetxea J., Agirre E., and Soroa A. Single or multiple? combining word representations independently learned from text and wordnet. *Proceedings of AAAI*, 2608–2614, 2016.

Goikoetxea J., Soroa A., Agirre E., and Donostia B.C. Random walks and neural network language models on knowledge bases. *Proceedings of HLT-NAACL*, 1434–1439, 2015.

Goldberg Y. and Levy O. word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

Gonzalez-Agirre A., Laparra E., and Rigau G. Multilingual central repository version 3.0. *LREC*, 2525–2529, 2012.

González Aguirre A. Computational models for semantic textual similarity. 2017.

Griffiths T.L., Steyvers M., and Tenenbaum J.B. Topics in semantic representation. *Psychological review*, 114(2):211, 2007.

Halawi G., Dror G., Gabrilovich E., and Koren Y. Large-scale learning of word relatedness with constraints. *Proceedings of the ACM SIGKDD*, 1406–1414, 2012.

Harris Z.S. Methods in structural linguistics. 1951.

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

Harris Z.S. Distributional structure. *Papers in structural and transformational linguistics*, 775–794. Springer, 1970.

Hassan S. and Mihalcea R. Cross-lingual semantic relatedness using encyclopedic knowledge. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, 1192–1201. Proceedings of ACL, 2009.

Hassan S. and Mihalcea R. Semantic relatedness using salient semantic analysis. *Aaai*. San Francisco, CA, 2011.

Hebb D.O. *The organization of behavior: A neuropsychological theory.* Psychology Press, 1949/2005.

Hill F., Cho K., Jean S., Devin C., and Bengio Y. Not all neural embeddings are born equal. *arXiv preprint arXiv:1410.0718*, 2014.

Hill F., Reichart R., and Korhonen A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

Hofmann T. Probabilistic latent semantic analysis. *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 289–296. Morgan Kaufmann Publishers Inc., 1999.

Hotelling H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Jauhar S.K., Dyer C., and Hovy E. Ontologically grounded multi-sense representation learning for semantic vector space models. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 683–693, 2015.

Kennedy A. and Hirst G. Measuring semantic relatedness across languages. *Proceedings, xLiTe: Cross-Lingual Technologies Workshop at the Neural Information Processing Systems Conference*, 2012.

Lazaridou A., Dinu G., and Baroni M. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1 lib., 270–280, 2015.

Lazic N., Subramanya A., Ringgaard M., and Pereira F. Plato: A selective context model for entity resolution. *Transactions of the Association for Computational Linguistics*, 3:503–515, 2015.

Leacock C. and Chodorow M. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998.

Leturia I. Evaluating different methods for automatically collecting large general corpora for basque from the web. *Proceedings of COLING*, 1553–1570, 2012.

Levy O. and Goldberg Y. Neural word embedding as implicit matrix factorization. *Proceedings of Advances in neural information processing systems*, 2177–2185, 2014.

Levy O., Goldberg Y., and Dagan I. Improving distributional similarity with lessons learned from word embeddings. *Proceedings of TACL*, 3:211–225, 2015.

Li Y., Zheng R., Tian T., Hu Z., Iyer R., and Sycara K. Joint embedding of hierarchical categories and entities for concept categorization and dataless classification. *arXiv preprint arXiv:1607.07956*, 2016.

Lin D. *et al.*. An information-theoretic definition of similarity. *Icml*, 98 lib., 296–304. Citeseer, 1998.

Liu Q., Jiang H., Wei S., Ling Z.H., and Hu Y. Learning semantic word embeddings based on ordinal knowledge constraints. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1 lib., 1501–1511, 2015.

Lopez-Gazpio I., Maritxalar M., Gonzalez-Agirre A., Rigau G., Uria L., and Agirre E. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems*, 119:186–199, 2017.

Lu A., Wang W., Bansal M., Gimpel K., and Livescu K. Deep multilingual correlation for improved word embeddings. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 250–256, 2015.

Luke S.G. and Christianson K. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 1–8, 2017.

Lund K. Semantic and associative priming in high-dimensional semantic space. *Proc. of the 17th Annual conferences of the Cognitive Science Society, 1995*, 1995.

Lund K. and Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208, 1996.

Mihalcea R. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 411–418. Association for Computational Linguistics, 2005.

Mikolov T., Chen K., Corrado G., and Dean J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Mikolov T., Le Q.V., and Sutskever I. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013b.

Mikolov T., Sutskever I., Chen K., Corrado G.S., and Dean J. Distributed representations of words and phrases and their compositionality. *Proceedings of Advances in neural information processing systems*, 3111–3119, 2013c.

Mikolov T., Yih W.t., and Zweig G. Linguistic regularities in continuous space word representations. *Proceedings of HLT-NAACL*, 746–751, 2013d.

Miller G.A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Miller G.A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Mrkšić N., Séaghdha D.O., Thomson B., Gašić M., Rojas-Barahona L., Su P.H., Vandyke D., Wen T.H., and Young S. Counter-fitting word vectors to linguistic constraints. *arXiv preprint arXiv:1603.00892*, 2016.

Mrkšić N., Vulić I., Séaghdha D.Ó., Leviant I., Reichart R., Gašić M., Korhonen A., and Young S. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints. *arXiv preprint arXiv:1706.00374*, 2017.

Navigli R. and Ponzetto S.P. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, 2012.

Neisser U. *Cognitive psychology: Classic edition.* Psychology Press, 1967/2014.

Nguyen K.A., Walde S.S.i., and Vu N.T. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. *arXiv preprint arXiv:1605.07766*, 2016.

Nickel M. and Kiela D. Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems*, 6341–6350, 2017.

Ono M., Miwa M., and Sasaki Y. Word embedding-based antonym detection using thesauri and distributional information. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 984–989, 2015.

Osborne D., Narayan S., and Cohen S.B. Encoding prior knowledge with eigenword embeddings. *arXiv preprint arXiv:1509.01007*, 2015.

Pennington J., Socher R., and Manning C.D. Glove: Global vectors for word representation. *Proceedings of EMNLP*, 14 lib., 1532–1543, 2014.

Piaget J. *Language and Thought of the Child: Selected Works vol 5.* Routledge, 1959/2005.

Pilehvar M.T., Jurgens D., and Navigli R. Align, disambiguate and walk: A unified approach for measuring semantic similarity. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1 lib., 1341–1351, 2013.

Pociello E., Agirre E., and Aldezabal I. Methodology and construction of the basque wordnet. *Language resources and evaluation*, 45(2):121–142, 2011.

Press W., Teukolsky S., Vetterling W., and Flannery B. *Numerical Recipes: The Art of Scientific Computing V 2.10 With Linux Or Single-Screen License.* Cambridge University Press, 2002.

Quine W.V.O. *Ontological relativity and other essays*. Number 1. Columbia University Press, 1969.

Radinsky K., Agichtein E., Gabrilovich E., and Markovitch S. A word at a time: computing word relatedness using temporal semantic analysis. *Proceedings of WWW*, 337–346, 2011.

Rastogi P., Van Durme B., and Arora R. Multiview LSA: Representation Learning via Generalized CCA. *Proceedings of HLT-NAACL*, 556–566, 2015.

Recski G., Iklódi E., Pajkossy K., and Kornai A. Measuring semantic similarity of words using concept networks. *Proceedings of the 1st Workshop on Representation Learning for NLP*, 193–200, 2016.

Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Rothe S. and Schütze H. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1793–1803, Beijing, China, 2015.

Roventini A., Alonge A., Bertagna F., Calzolari N., Cancila J., Girardi C., Magnini B., Marinelli R., Speranza M., and Zampolli A. Italwordnet: building a large semantic database for the automatic treatment of italian. *Linguistica Computazionale, Special Issue (XVIII-XIX)*, 745–791, 2003.

Rubenstein H. and Goodenough J.B. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

Ruder S., Vulić I., and Søgaard A. A Survey Of Cross-lingual Word Embedding Models. *ArXiv e-prints*, 2017.

Ruder S. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902, 2017.

Saussure F. *Course in general Linguistics*. Duckworth.(Translated by Roy Harris), 1916/1983.

Scharnhorst K. Angles in complex vector spaces. *Acta Applicandae Mathematica*, 69(1):95–103, 2001.

Schnabel T., Labutov I., Mimno D., and Joachims T. Evaluation methods for unsupervised word embeddings. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 298–307, 2015.

Schwartz R., Reichart R., and Rappoport A. Symmetric pattern based word embeddings for improved word similarity prediction. *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 258–267, 2015.

Shanks D.R. Learning: From association to cognition. *Annual review of psychology*, 61:273–301, 2010.

Sinha R. and Mihalcea R. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. *Semantic Computing, 2007. ICSC 2007. International Conference on*, 363–369. IEEE, 2007.

Skinner B.F. *Verbal behavior*. BF Skinner Foundation, 1957/2014.

Socher R., Lin C.C., Manning C., and Ng A.Y. Parsing natural scenes and natural language with recursive neural networks. *Proceedings of the 28th international conference on machine learning (ICML-11)*, 129–136, 2011.

Spearman C. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Speer R. and Havasi C. Conceptnet 5: A large semantic network for relational knowledge. *The PeopleâĂŹs Web Meets NLP*, 161–176. Springer, 2013.

Tian F., Gao B., Chen E., and Liu T.Y. Learning better word embedding by asymmetric low-rank projection of knowledge graph. *arXiv preprint arXiv:1505.04891*, 2015.

Turian J., Ratinov L., and Bengio Y. Word representations: a simple and general method for semi-supervised learning. *Proceedings of the 48th annual meeting of the association for computational linguistics*, 384–394. Association for Computational Linguistics, 2010.

Tversky A. Features of similarity. *Psychological review*, 84(4):327, 1977.

Vosniadou S. and Ortony A. *Similarity and analogical reasoning*. Cambridge University Press, 1989.

Vossen P., editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Springer Netherlands, 1998.

Vygotsky L.S. *Mind in society: The development of higher psychological processes*. Harvard university press, 1980.

Wang H., Raj B., and Xing E.P. On the origin of deep learning. *arXiv preprint arXiv:1702.07800*, 2017.

Wang Z., Zhang J., Feng J., and Chen Z. Knowledge graph embedding by translating on hyperplanes. *Proceedings of AAAI*, 1112–1119. Citeseer, 2014.

Weaver W. Translation. *Machine translation of languages*, 14:15–23, 1955.

Wick M., Kanani P., and Pocock A. Minimally-constrained multilingual embeddings via artificial code-switching. *Proceedings of AAAI*, 2849–2855, 2016.

Wieting J., Bansal M., Gimpel K., Livescu K., and Roth D. From paraphrase database to compositional paraphrase model and back. 2015.

Wittek P., Koopman B., Zuccon G., and Darányi S. Combining word semantics within complex hilbert space for information retrieval. *International Symposium on Quantum Interaction*, 160–171. Springer, 2013.

Witten I.H. and Milne D.N. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. 2008.

Wittgenstein L. *Tractatus logico-philosophicus*. Routledge, 1921/2013.

Wittgenstein L. Philosophische untersuchungen i philosophical investigations (gem. anscombe & r. reesh, eds.,), 1953.

Xing C., Wang D., Liu C., and Lin Y. Normalized word embedding and orthogonal transform for bilingual word translation. *Proceedings of HLT-NAACL*, 1006–1011, 2015.

Xu C., Bai Y., Bian J., Gao B., Wang G., Liu X., and Liu T.Y. Rc-net: A general framework for incorporating knowledge into word representations. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 1219–1228. ACM, 2014.

Xu H., Murphy B., and Fyshe A. Brainbench: A brain-image test suite for distributional semantic models. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2017–2021, 2016.

Yu M. and Dredze M. Improving lexical embeddings with semantic knowledge. *ACL (2)*, 545–550, 2014.

Zhang Y., Gaddy D., Barzilay R., and Jaakkola T. Ten pairs to tag–multilingual pos tagging via coarse mapping between embeddings. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1307–1317, 2016.

Glosategia

adiera (*sense*)

Hitz baten esanahi anitzetako bakoitza.

agerkidetza (*co-ocurrence*)

Dokumentu batean termino bi edo gehiagok elkarrekin agertzea, zorizkoa baino maiztasun handiagoarekin.

ahaidetasun semantiko (*semantic relatedness*)

Antzekotasuna baino kontzeptu orokorragoa, sinonimia eta hiperonimia erlazioez gain, meronimia (oin/behatz), antonimia (argi/ilun), asoziazio funtzionalak (arropa/armairu) eta bestelako ezohiko erlazioak bere barne hartzen dituena.

antzekotasun- eta ahaidetasun-emaitzen batezbestekoen konbinaketa, BB

Testu-corpusetako eta ezagutza-baseetako informazio semantikoa uztartzeko metodoa, bi iturri horien antzekotasun-emaitzen batezbestekoak kalkulatuta lortzen dena.

antzekotasun semantiko(*semantic similarity*)

Hitzen arteko sinonimia (okela/haragi), hiponimia/hiperonimia (kolore/urdin) erlazioak.

aurreranzko propagazio (*forward-propagation*)

Neurona-sareetan sarrerak geruzetan zehar duen propagazioa, irteera bat sortzen duena.

ausazko ibilbide (*random walk*)

Ezagutza-base baten informazio estrukturala ustiatzeko metodo globala. Ezagutza-basea grafo legez ulertuta, ibilbideak grafoko ertzez erlazionatutako erpinak zeharkatuz burutzen dira. Ausazko ibilbideek batek erpinen garrantzia estruktural erlatiboa neurtzen dute.

atzeranzko propagazio (*back-propagation*)

Neurona-sare artifizialetako pisuak gradienteen bidez eguneratzeko metodoa. Aurreranzko propagazioa (ik. *aurreranzko propagazio*) burutu ondoren, galera-funtzioaren gradienteekin neurona-sareko pisuak eguneratzen dira.

azpi-laginketa atalase (*subsampling threshold*)

Word2vec ereduaren parametroa, maiztasun handiko hitzak ausaz hiztegitik ezabatzeko probabilitatea kalkulatzeko erabilia.

bateratze-metodo (*joint method*)

Ikasketa-prozesuan testu corpusetako eta ezagutza-baseetako informazioa konbinatzen dituen eredua. Sarreran corpusak eta ezagutza-baseetako erlazio-zerrendak dituzte, hitzen errepresentazioak hasieratik kalkulatuz.

bektoreen kateaketa, KAT

Testu-corpusetako eta ezagutza-baseetako informazio semantikoa uztartzeko metodoa, testutik eta ezagutza-basetik erauzitako hitzen errepresentazioak kateatuz lortzen dena.

Continuous-Bag-of-Words, CBOW

Word2vec (ik. *word2vec*) eredu-multzoko eredua. Testuinguruko hitzen esanahietatik abiatuta, behatutako hitzaren esanahia aurresaten du.

CBOW corpus sintetikoen gainean, AISG

Ezagutza-base batetik corpus sintetikoa (ik. *corpus sintetiko*) erauzita, azken hori CBOWekin (ik. *CBOW*) prozesatu ondoren kalkulatutako errepresentazio trinkoak. Errepresentazio horiek ezagutza-baseko informazio estrukturala jasotzen dute.

corpusen konbinaketa, HIB

Testu-corpusetako eta ezagutza-baseetako informazio semantikoa uztartzeko metodoa, ik. *corpus hibrido*.

corpus sintetiko, CS (*synthetic corpus*)

Ezagutza-baseen gainean ausazko ibilbideen metodo batekin erauzitako corpusa. Ibilbide horietan zeharkatutako kontzeptuen lexikalizazioak ausaz aukeratzen dira, eta fitxategi batean idazten. Corpus sintetikoek ezagutza-baseko informazio estrukturala inplizituki kodetzen dute, agerkidetzen bidez. Elebakarrak zein eleaniztunak izan daitezke.

corpus hibrido (*hybrid corpus*)

Testu corpus eta corpus sintetikoen konbinaketa. Corpus hibridoek testu corpusetako eta ezagutza-baseetako agerkidetzak nahasten dituzte, bien informazio semantikoa konbinatuz. Elebakarrak zein eleaniztunak izan daitezke.

datu-multzo (*dataset*)

Hitzen errepresentazioen ebaluazioa egiteko datu-bilduma, hitz-bikoteez eta azken horien giza antzekotasun-balioez osatua.

dimentsionaltasun (*dimensionality*)

Bektore-espazio baten dimentsio kopurua.

dump

Ezagutza-base baten bertsio jakin baten egitura eta edukiaren *backupa*. Tesi-lan honetan Wikipediakoak erabili ditugu.

elearteko antzekotasun (*cross-lingual similarity*)

Ele desberdinetako hitzen arteko antzekotasun semantikoa.

erpin (*vertex*)

Grafo bat osatzen duen oinarrizko elementua. Erpinak eurena artean lotuta daude, ertzen bitartez.

erro-bilatzaile (*stemmer*)

Hitzen erroak bilatzeko tresna. Erregela finko batzuez baliatuta, hitzei atzizki ohikoenak kentzen dizkie. Ez du baliabide edo prozesu linguistikorik erabiltzen.

ertz (*edge*)

Grafo bateko erpinak lotzen dituen erlazioa. Grafoetako ertzak norabidea badute ertz zuzenduak dira, eta bestela ez-zuzenduak. Gainera, ertzek pisuren bat esleituta izan dezakete.

eredu semantiko distribuzional, ESD (*distributional semantic model*)

Semantika Distribuzionalean oinarritutako eredu konputazionala. Hitzen errepresentazio distribuzionalak bektore-espazio batean kodetzen dituzte.

esangura-test (*significance-test*)

Estatistikan, emaitza bat estatistikoki esanguratsua dela jakiteko testa.

esangura-maila (*significance level*)

Estatistikan, emaitza bat estatistikoki esanguratsua dela jakiteko probabilitate-atalasea.

estatistikoki esanguratsu (*statistically significant*)

Estatistikan, emaitza jakin batek ausaz gertatzeko aukerarik ez duenean.

ezagutza-base (*knowledge base*)

Kontzeptuei buruzko informazio egituratua duen biltegia edo lexikoia.

ezagutza-baseko errepresentazio trinko

Bektore eskalar trinkoa bat da, eta ezagutza-baseko hitz baten esanahia errepresentatzen du. Bektore horiek corpus sintetikoetako (ik. *corpus*

sintetiko) hitz-testuinguru agerkidetzekin kalkulatzen dira, eta beren dimentsio bakoitza esanahiaren tasun semantiko latente bat da.

fintze-metodo (*inject method*)

Aurre-entrenatutako hitz-bektoreak ezagutza-baseetako erlazioen bidez fintzeko metodoa.

galera-funtzio (*loss function*)

Optimizazio problemetan minimizatu (edo, batzuetan, maximizatu) beharreko funtzioa. Galera-funtzioak irteera legez zenbaki erreal bat kalkulatzen du, optimizazioaren kostua adierazten duena, eta, ondorioz, txikitzen joan behar dena. Optimizazioa, oro har, parametroen estimaziorako erabiltzen da.

gradiente-jaitsiera estokastikoa (*stochastic gradient descent*)

Galera-funtzio bat minimizatzeko gradienteetan oinarritutako algoritmoa estokastikoa. Optimizazio problemako parametroak eguneratzeko erabiltzen da.

grafo (*graph*)

Matematiketan, eta bereziki grafoen teorian, elkarren artean erlazionatutako objektuen multzo egituratuari egiten dio erreferentzia. Objektuak "erpin" deituriko abstrakzio matematikoak dira, eta azken horien arteko erlazioak "ertz" bezala ezagutzen dira.

hitz-bektore (*embedding*)

Bektore eskalar trinko bat da, eta hitz baten esanahia errepresentatzen du. Bektore horiek testu corpusetako hitz-testuinguru agerkidetzekin kalkulatzen dira, eta beren dimentsio bakoitza esanahiaren tasun semantiko latente bat da.

hizkuntza-eredu (*language model*)

Hitz-sekuentzien probabilitateak kalkulatzen dituen eredua probabilistikoa. N-gramen baitako hitz-sekuentzia bat jakinik, hurrengo hitza aurresaten dute. Hizkuntzaren prozesamenduan oso ospetsuak dira, eta egungo neurona-sareetan oinarritutako iragarpen-metodoen muinean daude.

hizkuntzaren prozesamendu (*natural language processing*)

Hizkuntzaren tratamendu automatikoaren inguruko ikerketa-lerroa.

iragarpen-metodo (*predict methods*)

Testu corpusetako hitz-bektoreak kalkulatzeko eredu semantiko distribuzional gainbegiratua. Metodo horiek hizkuntza-ereduak eta neurona-sareak uztartzen dituzte.

isomorfismo (*isomorphism*)

Bi objektu matematikok egitura bera dutenean. Hizkuntzaren prozesamenduan, ele desberdinetako bektore-espazio bereiziek egiturari egiten die erreferentzia.

kontaketa-metodo (*count method*)

Testu corpusetako hitz-bektoreak kalkulatzeko eredu semantiko distribuzional ez-gainbegiratua. Metodo horiek hitz-testuinguru kontaketak matrizeetan oinarritzen dira, baina kontaketak ez dira bere horretan uzten: kontaketei haztapen bat aplikatzen zaie, eta matrizearen dimentsio-murrizketa egiten zaio.

kontzeptu (*concept*)

Errealitateko elementuen irudikapen abstraktuak.

korrelazio kanonikoen analisia, KKA (*canonical correlation analysis*)

Ausazko aldagaiez osatutako bi bektoretatik abiatuta, eta aldagaien artean korrelazioa egonik, metodo honek bektoreen korrelazio maximoa lortzeko konbinaketa lineala kalkulatzen du. Tesi-lan honetan testu-corpusetako eta ezagutza-baseetako informazio semantikoa uztartzeko metodoetako bat izendatzeko ere erabili da. Bada, konbinaketa-metodo horrek aipatutako bi iturrien bektore-espazioen arteko korrelazio maximoa duen espazioa partekatua kalkulatzen du, eta gero ezagutza-baseko errepresentazioak espazio komun horretan proiektatzen ditu.

kosinu-antzekotasun (*cosine similarity*)

Bi bektoreen antzekotasuna neurtzeko metodoa, biderkadura eskalarra erabilita. Antzekotasuna bektoreen angeluen arteko kosinua legez ulertzen da; 1 balioak hitzek esanahi bera dutela esan gura du, 0 balioak antzekotasunik ez dutela, eta -1 balioak guztiz aurkako esanahia dutela.

laginketa negatibo (*negative sampling*)

Word2vec tresnaren galera-funtzioaren optimizatzeko metodo eraginkorra. Laginketa negatiboa *softmax* funtzioaren hurbilpen eraginkorra, eta behatutako hitz-testuinguru pare bakoitzerako, ausazko hitz-testuinguru laginak soilik erabiltzen ditu. Neurona-sareko pisu guztiak ez dituenez eguneratzen, ikasketa-prozesua asko azkartzen du.

lexikalizazio (*lexicalization*)

ik. *variant*.

mapaketa lineal, mapaketa-metodo, map (*lineal mapping*)

Bi bektore-espazio elebakarren arteko transformazio lineala, azken horien arteko isomorfismoa mantentzen duena.

moteltze-faktore (*damping factor*)

PageRank algoritmoan ausazko ibilbideekin jarraitzeko ala geratzeko probabilitate-atalasea definitzen du.

neurona-sare (*neural network*)

Neurona biologikoen ikasketan inspiratutako eredu matematikoak. Aurretiaz inongo ezagutzarik eduki gabe, adibideen bidez ikasten dute. Hizkuntzaren prozesamenduan hizkuntza-ereduak neurona-sareen arkitekturan txertatu dira, eta testu corpusetako hitzen errepresentazioak kalkulatzen dituzte.

oinarri-lerro (*baseline*)

Ataza jakin baten oinarrizko soluzioa. Sistema berri batek esperimentazio-garaian oinarri-lerroaren emaitzak gainditu beharko ditu.

osagai nagusien analisia, ONA (*principal component analysis*)

Jatorrizko aldagai korrelaziodunak korrelaziorik gabeko beste aldagai kopuru txikiagora murrizten dituen metodoa, osagai deiturikoak. Osagaiak jatorrizko aldagaien transformazio linealaren bidez kalkulatzen

dira, eta azken horiek koordenatu sistema berri batean jartzen ditu; lehenengo ardatzak jatorrizko aldagaien artean bariantza handiena du, bigarren ardatza bigarrena da bariantzan, eta, modu horretan, bata bestearen ondoren. Tesi-lan honetan testu-corpusetako eta ezagutza-baseetako informazio semantikoa uztartzeko metodoetako bat izendatzeko ere erabili da. Bada, aipatutako bi iturrietako hitzen errepresentazioak kateatu ondoren (ik. *bektoreen kateaketa*) ONA aplikatzen zaie, eta dimentsionaltasun (ik. *dimentsionaltasun*) murritzagoko errepresentazio hibridoak sortzen ditu.

Personalized Pagerank

Grafoaren informazio estrukturala bere osotasunean ustiatzen duen algoritmoa, ausazko ibilbideetan oinarritutakoa. Grafoaren erpinen gainean ibilbideak egin aurretik, erpinei beren garrantzia estrukturalaren araberako pisua bat esleitzen die.

Personalized Pagerank bektoreak, PPB

Personalized PageRank (ik. *Personalized Pagerank*) algoritmoan oinarrituta, grafo batetik erauzitako hitzen errepresentazioak. Bektore horiek dimentsionaltasun (ik. *dimentsionaltasun*) handia daukate (grafoaren nodoak bestekoa), eta bere balioak oso sakabanatuta daude.

sailkapen-balioetan oinarritutako konbinazioa, RNK

Testu-corpusetako eta ezagutza-baseetako informazio semantikoa uztartzeko metodoa, bi iturri horien antzekotasun-emaitzen sailkapen-balioen (*ranking*) batezbestekoa kalkulatuta lortzen dena.

Skip-gram

Word2vec (ik. *word2vec*) eredu-multzoko eredua. Behatutako hitzaren esanahitik abiatuta, testuinguruko hitzenak aurresaten ditu.

Skip-gram corpus sintetikoen gainean, AISG

Ezagutza-base batetik corpus sintetikoa (ik. *corpus sintetiko*) erauzita, azken hori Skip-gramekin (ik. *Skip-gram*) prozesatu ondoren kalkulatutako errepresentazio trinkoak. Errepresentazio horiek ezagutza-baseko informazio estrukturala jasotzen dute.

Skip-gram corpus elebidunetan, BAT

Corpus elebidun bat osatu ondoren (testu-corpusa, corpus sintetikoa (ik. *corpus sintetiko*) edo corpus hibridoa (ik. *corpus hibrido*)) Skip-gram ereduarekin (ik. *Skip-gram*) kalkulatutako hitzen errepresentazioak.

Skip-gram murriztapenekin corpus elebidunetan, BATM

Corpus elebidun bat osatu ondoren (testu-corpusa, corpus sintetikoa edo corpus hibridoa) murriztapenekin hedatutako Skip-gram ereduarekin kalkulatutako hitzen errepresentazioak. Skip-gram hedatua bateratze-metodoa (ik. *bateratze-metodo*) da.

softmax

Ikasketa automatikoan erregresio logistikoaren orokortzea da, eta sailkapenerako erabiltzen da. Neurona-sareetan erabiliz gero, azken horren irteeran joaten da.

synset

WordNeten, kontzeptu lexikal edo adiera bati dagokion sinonimo-multzoa (*synonym set*). Synseta *variantek* osatzen dute.

testuinguru sintetiko (*synthetic context*)

Ezagutza-base batean, erlazionatutako kontzeptuez osatutako testuinguruak. Azken horiek ausazko ibilbideekin sortzen dira, eta ezagutza-baseko informazio estrukturala modu inplizituan gordetzen dute.

urre-patroi (*gold standard*)

Gizakien irizpideekin osatutako datu-multzoa (ik. *datu-multzo*). Eredu konputazionalen emaitzekin konparatzen dira, eta azken horiek ebaluatzeko balio dute.

variant

WordNet-en, synset bati esleitutako ale lexikaletako bakoitza. Esaterako, *moon*, *luna* eta *ilargi synset* bereko *variantak* dira, kontzeptu bera adierazten baitute.

weighted overlap, WO

Ezagutza-baseen adieren probabilitate-distribuzioekin osatutako bektoreen artean antzekotasuna kalkulatzeko metodoa. Bi bektoretan teilakatutako adieren sailkapenean oinarritzen da.

word2vec

Testu corpusetan oinarritutako iragarpen-metodoa (ik. *iragarpen-metodo*), hitzen esanahia bektore-espazio trinko batean kodetzen duena. Oso optimizatuta dago; bi geruza ditu soilik eta *softmaxa* barik laginketa negatiboa (ik. *laginketa negatiboa*) darabil. Bi eredu erabili ditzake, CBOW (ik. *CBOW*) eta Skip-gram (ik. *Skip-gram*).

WordNet

Hitz eta adierei buruzko informazioa duen ezagutza-base lexikala, jatorrian ingelesaren gainean osatua, baina egun hainbat eletara hedatua. Izenez, aditzez, adjektiboz eta adberbioz osatutako egitura da, synset delakoen arabera antolatuta dago eta azken horiek hainbat erlazio semantikorekin lotuta daude.

zenbaki konplexuen konbinaketa, KNP

Testu-corpusetako eta ezagutza-baseetako informazio semantikoa uztartzeko metodoa, testutik eta ezagutza-basetik erauzitako hitzen errepresentazioak zenbaki konplexu baten formatuan jarrita lortzen dena. Testutik erauzitako errepresentazioa zati errealean jartzen da, eta ezagutza-basekoa konplexuan.

A Eranskina

A.1 Skip-gram eredua

Eranskin honen helburua 3. kapituluan azaldutako Skip-gram ereduaren xehetasun gehiago ematea da. Lehenik, Skip-gramen galera-funtzioa, bere gradienteak eta azken horiei dagozkien aldagaien eguneraketak jarriko ditugu, eta, ondoren, Skip-gram kodearekin lotuko ditugu. Gainera, kodean aurreranzko propagazioaren eta atzeranzko propagazioaren artean desberdinduko dugu. Kode hori `word2vec`[1] algoritmoaren zati bat da, eta C lengoaian idatzita dago.

Bada, (A.1) ekuazioa[2] Skip-gram ereduk galera-funtzioa da. (A.2), (A.3) eta (A.4) ekuazioak galera-funtzio horren aldagaiekiko gradienteak dira, hots, c, w eta w_n aldagaiekiko (behatutako testuinguruekiko, behatutako hitzekiko eta lagin negatiboekiko, hurrenez hurren). Kontuan izan, (A.1) ekuazioko galera-funtzioa leiho baten baitako (w, c) behatutako hitz-testuinguru bikote batentzako dela soilik, eta gradieten horiek atzerantzko propagazioan burutzen direla.

$$J_{sg}(w, c) = log(\sigma(w^T c)) + \sum_{n=1}^{N} \mathbb{E}_{w_n \sim P(w)} \Big[log(\sigma(-w_n^T c)) \Big] \tag{A.1}$$

[1] Lan honetan jatorrizko kodea erabiliko dugu, eta azken hori `https://code.google.com/archive/p/word2vec/` estekan dago. Egun hurrengo `github` errepositorioan ere aurki daiteke: `https://github.com/dav/word2vec`

[2] 3. kapituluko (3.3) ekuazioa da, errepikatua.

$$\frac{\partial J_{sg}}{\partial c} = (1 - \sigma(w^t c))w - \sum_{n=1}^{N}(1 - \sigma(-w_n^t c))w_n \tag{A.2}$$

$$\frac{\partial J_{sg}}{\partial w} = (1 - \sigma(w^t c))c \tag{A.3}$$

$$\frac{\partial J_{sg}}{\partial w_n} = -(1 - \sigma(-w_n^t c))c \tag{A.4}$$

Bada, (A.2) eta (A.3) ekuazioko gradienteak behin kalkulatuko dira, eta (A.4) ekuaziokoa N lagin negatibo guztientzat. Atzeranzko propagazioaren hurrengo pausuan gradiente horiekin agerkidetzetako hitz-bektoreak eguneratzen dira. (A.5), (A.6) eta (A.7) ekuazioek c, w eta w_n hitz-bektoreen eguneraketak deskribatzen dituzte, hurrenez hurren:

$$c = c + \mu\frac{\partial J_{sg}}{\partial c} \tag{A.5}$$

$$w = w + \mu\frac{\partial J_{sg}}{\partial w} \tag{A.6}$$

$$w_n = w_n + \mu\frac{\partial J_{sg}}{\partial w_n} \tag{A.7}$$

Ekuazio horietan μ Skip-gram ereduaren ikasketa-indizea da. Eguneraketa kopurua gradiente kopuruen bera da.

Listing A.1 – Skip-gram ereduaren ikasketa-prozesuko muinaren kodea.

```
for (a = b; a < window * 2 + 1 - b; a++) if (a != window) {
   c = sentence_position - window + a;
   if (c < 0) continue;
   if (c >= sentence_length) continue;
   last_word = sen[c];
   if (last_word == -1) continue;
   l1 = last_word * layer1_size;
   for (c = 0; c < layer1_size; c++) neu1e[c] = 0;
   if (negative > 0) for (d=0; d<negative+1; d++) {
    if (d == 0) {
            target = word;
            label = 1;
        } else {
            next_random = next_random * (unsigned long long)25214903917+11;
            target = table[(next_random >> 16) % table_size];
            if (target == 0) target = next_random % (vocab_size - 1) + 1;
            if (target == word) continue;
            label = 0;
        }
        l2 = target * layer1_size;
        f = 0;
        for (c = 0; c<layer1_size; c++) f += syn0[c + l1] * syn1neg[c + l2];
        if (f > MAX_EXP) g = (label - 1) * alpha;
        else if (f < -MAX_EXP) g = (label - 0) * alpha;
        else g=(label-expTable[(int)((f+MAX)*(EXP_SIZE/MAX/2))])*alpha;
        for (c = 0; c<layer1_size; c++) neu1e[c] += g * syn1neg[c + l2];
        for (c = 0; c<layer1_size; c++) syn1neg[c + l2] += g * syn0[c + l1];
   }
   for (c = 0; c < layer1_size; c++) syn0[c + l1] += neu1e[c];
}
```

A.1 kodea Skip-gram ereduko ikasketaren muina da. 3 kapituluan aipatu dugunez, W hitzen espazioa eta C testuinguruen espazioa `syn1neg` eta `syn0` legez izendatzen dira kodean, hurrenez hurren. Hurrengo puntuetan kodeko aldagaiak eranskin honetan aipatutako ekuaziokoekin lotuko ditugu:

- `window`[3]: testuinguru-leihoaren zabalera.
- `negative`[4]: lagin negatibo kopurua.
- `d`: behatutako w hitzaren edo w_n lagin negatiboaren artean desberdintzeko aldagaia:
  - `d`=0: behatutako w-rekin lan egin.
  - 0<`d`<`negative`: w_n lagin negatiboekin lan egin.

[3](A.1) ekuazioa (w, c) agerkidetza bakarrerako da, eta parametro hori ez da agertzen.
[4](A.1) eta (A.2) ekuazioetako N.

- `last_word`: behatutako c testuinguruaren indizea hiztegian.
- `word`: behatutako w hitzaren indizea hiztegian.
- `target`: behatutako w hitzaren edo w_n lagin negatiboaren indizea hiztegian:
  - `d`=0: behatutako w, `target`=`word`.
  - 0<`d`<`negative`: w_n lagin negatiboa, ausazko indizea `target`-en.
- `l1`: behatutako c testuinguruaren indizea `syn0` espazioan, `last_word`-ekin kalkulatua.
- `l2`: behatutako w hitzaren (`d`=0) edo w_n lagin negatiboaren (0<`d`<`negative`) indizea `syn1neg` espazioan, `target`-ekin kalkulatua.
- `g`: c, w eta w_n gradienteekin eguneratzeko bitarteko terminoa. Xehetasun gehiago kodea lerroz lerro azaltzerakoan.
- `alpha`[5]: ikasketa-indizea.
- `syn1neg[c + l2]`[6]: behatutako w hitzaren (`d`=0), edo w_n lagin negatiboaren (0<`d`<`negative`) hitz-bektoreak.
- `syn0[c + l1]`[7]: behatutako c testuinguruaren hitz-bektorea.

Hurrengo puntuetan A.1 kodean aurrerantzko propagazioaren eta atzerantzko propagazioaren arteko bereziketa egin dugu, eta lerro esanguratsuenen edukiak laburbildu ditugu:

- 1. lerroa: `for` begizta nagusia, 2×`window`+1 iteraziotakoa, (w, c) hitz-testuinguru agerkidetza posible guztiak osatzeko erabilia. Ez da (A.1) ekuazioan agertzen.
- 7. lerroa: `l1` eguneratu behatutako c-ren `syn0` espazioko indizearekin.
- 9-27. lerroak: agerkidetzen `for` begizta, `negative`+1 iteraziotakoa[8]:
  - 10-19 lerroak: `target` w edo w_n aldagaien hiztegiko indizearekin eguneratu.
  - 20. lerroa: `l2` w edo w_n aldagaien `syn1neg` espazioko indizeekin eguneratu.

[5] (A.5), (A.6) eta (A.7) ekuazioetako μ.

[6] Ataleko honetako ekuaziokoetako w edo w_n.

[7] Ataleko honetako ekuaziokoetako c.

[8] Hurrengo lerroetan, `d`=0 denean (w, c) agerkidetzarekin lan egingo dugu, eta gainontzekoetan (w_n, c) agerkidetzekin.

  - Aurrerantzko propagazioa:
    * 22. lerroa: `f` aldagaiak (w, c) edo (w_n, c) agerkidetzen biderketa da.
  - Atzerantzko propagazioa:
    * 23-25. lerroak: `f` aldagaiaren sigmoidea[9] kalkulatu ondoren, `g` kalkulatu:
      · (w, c) denean: c eta w aldagaiekiko gradienteak kalkulatzeko bitarteko terminoak[10].
      · (w_n, c) denean: w_n aldagaiarekiko gradienteak kalkulatzeko bitarteko terminoa[11].
    * 26. lerroa: c gradientea akumulatu[12] `neu1e` aldagaian.
    * 27. lerroa: w eta w_n hitz-bektoreak eguneratu[13].
- 29. lerroa: c hitz-bektorea eguneratu[14].

A.2 Murriztapenak Skip-gram ereduan

Eranskin honen helburua 5. kapituluan aurkeztutako murriztapenekin hedatutako Skip-gram ereduaren xehetasun gehiago ematea da. Lehenik, Skip-gram headtuaren galera-funtzioa, bere gradienteak eta azken horiei dagozkien aldagaien eguneraketak jarriko ditugu, eta, ondoren, inplementazio horren kodearekin lotuko ditugu. Ikusiko dugunx, kodea hori atzeranzko propagazioan murriztapenek egindako ekarpenei soilik dagokie, gainontzekoa bere horretan geratzen baita (ik. A.1 eranskina). Kode hori `word2vec`[15] algoritmoaren gainean inplementatu dugu, eta C lengoaian idatzita dago.

[9] (A.1) ekuazioko $\sigma(w^T c)$ eta $\sigma(w_n^T c)$ taula baten bitartez kalkulatzen du. Ikusi `https://goo.gl/3DBUaz` esteka.

[10] (A.2) eta (A.3) ekuazioak μ aldagaiagatik biderkatuz gero, $\mu * (1 - \sigma(w^T c))$ adierazten dute.

[11] (A.4) ekuazioa μ aldagaiagatik biderkatuz gero, $-\mu * (1 - \sigma(w_n^T c))$ adierazten du. Kodeari so, $\mu * (0 - \sigma(w_n^T c))$ da, baina $\sigma(-x) = 1 - \sigma(x)$ baliokidetza erabiltzen dute.

[12] (A.2) ekuazioa. `d`=0 denean ekuazioaren ezkerraldeko termino, 0<`d`<`negative` denean eskubiko batukaria.

[13] (A.6) eta (A.7) ekuazioak.

[14] (A.5) ekuazioa.

[15] Lan honetan jatorrizko kodea erabiliko dugu, eta azken hori `https://code.google.com/archive/p/word2vec/` estekan dago. Egun hurrengo `github` errepositorioan ere aurki daiteke: `https://github.com/dav/word2vec`

Listing A.2 – Murriztapenen kodea Skip-gram ereduan.

```
    for (s = 0; s < vocab[target].sim_constrnum; s++){
      s1 = sim[vocab[target].sim_indx].constr_array[s] * layer1_size;
      if (weightsim == 1){
        if(vocab[target].sim_constrnum>1){
          relative_weight=(real)sim[vocab[target].sim_indx].weight_array[s];
        }
      }else{
          relative_weight = 1;
      }
      for (c = 0; c < layer1_size; c++){
        grad[c] += 2*lambdasim*(syn1neg[c + target*layer1_size] \
                                - relative_weight*syn1neg[c + s1]);
        syn1neg[c + s1] += alpha*2*lambdasim*relative_weight \
                           *(syn1neg[c + target*layer1_size] \
                             - relative_weight*syn1neg[c + s1]);
      }
    }
```

Bada, (A.8) ekuazioan[16] oinarritutakoa.

- (A.9) ekuazioa[17] w (behatutako hitza) aldagaiarekiko gradiente hedatua da.
- (A.10) ekuazioa[18] w_{lm} (behatutako hitzaren murriztapen eleanitzak) aldagaiarekiko gradiente berria da.
- c eta w_n aldagaiekiko (behatutako testuinguruekiko eta lagin negatiboekiko, hurrenez hurren) gradienteak[19] bere horretan geratzen dira.

$$J_{sg+}(w,c) = J_{sg}(w,c) - \lambda \sum_{l=1}^{2} \sum_{m=1}^{M_l(w)} ||w - w_{lm}||_2^2 \tag{A.8}$$

$$\frac{\partial J_{sg+}}{\partial w} = (1 - \sigma(w^t c))c - 2\lambda \sum_{l=1}^{2} \sum_{m=1}^{M_l(w)} (w - w_{lm}) \tag{A.9}$$

[16] 5.2.1. ataleko (5.1) ekuazioa, errepikatua.
[17] 5.2.1. ataleko (5.2) ekuazioa, errepikatuta.
[18] 5.2.1. ataleko (5.3) ekuazioa, errepikatuta.
[19] (A.2) eta (A.4) ekuazioak, hurrenez hurren.

$$\frac{\partial J_{sg+}}{\partial w_{lm}} = 2\lambda \sum_{l=1}^{2} \sum_{m=1}^{M_l(w)} (w - w_{lm}) \tag{A.10}$$

A.2 kodearen murriztapenen gradienteak kalkulatzeko kodea da, eta A.1 kodearen 27. lerroan txertatu dugu, azpiprograma batetan.

Kontuan izan, A.2 kodea A.1 kodearen agerkidetzen `for` begiztaren barruan (8-28 lerroak) dagoela. Hurrengo puntuek A.2 kodeko aldagai eta parametroak (A.8), (A.9) eta (A.10) ekuazioetakoekin lotzen dituzte:

- `sim_constrnum`[20]: w behatutako hitzaren murriztapen kopurua. Inplementazio honetan hizkuntza guztietako murrizketak batera doaz.
- `s1`: w_{lm} murriztapenak `syn1neg` espazioko indizea.
- `wn_indx`: w behatutako hitzaren WordNet kontzeptuen zerrendako (guk sortutakoa) indizea. Ez bada WordNet-en existitzen, -1 balio du.
- `weightsim`: w behatutako hitzaren lexikalizazio-murriztapenekin edo synset-murriztapenekin lan egitea ahalbidetu[21].
- `weight_array`: w behatutako hitzaren synset bakoitzari pisuak esleitzeko array-a[22].
- `relative_weight`: w behatutako hitzaren synset-murriztapen jakin baten pisuaren balioa[23].
- `lambda_sim`[24]: erregularizazio terminoa.
- `syn1neg[c + s1]`[25]: w_{lm} murriztapenaren hitz-bektorea.

w behatutako hitzari A.2 kodeko murriztapenen kalkuluak behin bakarrik aplikatzen zaizkio. Zehazki, (w, c) behatutako hitz-testuinguru agerkidetzako kalkuluekin[26] batera egiten dira, (w_n, c) agerkidetzen aurretik. Gainera,

[20] (A.8) ekuazioari so, l*$M_l(w)$ balio du.
[21] Tesi-lan honetan ez dugu synset-murriztapenik erabili (`weightsim`=0).
[22] Tesi-lan honetan ez dugu synset-murriztapenik erabili.
[23] Tesi-lan honetan ez dugu synset-murriztapenik erabili (`relative_weight`=1)
[24] (A.8), (A.9) eta (A.10) ekuazioetako λ.
[25] (A.8), (A.9) eta (A.10) ekuazioetako w_{lm}.
[26] A.1 kodean `d`=0 denean.

A.2 kodea WordNet-en existitzen diren lexikalizazioei soilik aplikatzen zaie (`wn_indx`≠-1). Hurrengo puntuek kode horren lerro esanguratsuenen edukiak laburbiltzen dituzte:

- 1. lerroa: Murriztepenen `for` begizta, `sim_constrnum` iteraziotakoa.
- 2. lerroa: `s1` indizearen kalkulua.
- 3-9. lerroak: murriztapenen pisuen kudeaketa. Defektuz, ez zaie pisurik esleitzen (`weightsim`=0 eta `relative_weight`=1).
- 11-12. lerroa: gradientea akumulatzen du `grad` aldagaian, murriztapen guztiekin[27].
- 13-15. lerroak: w_{lm} murriztapenaren hitz-bektorearen eguneraketa[28].

A.1 kodeko `window` zabalerako leihoko kalkulu guztiak amaitzen dituenean[29] behatutako w hitz-bektorea[30] eguneratzen du. (A.3) ekuazioak eguneraketa horren kodea deskribatzen du:

Listing A.3 – w hitzaren eguneraketa gehigarria, murriztapenak kontuan hartuta.

```
1 if(vocab[word].wn_indx != -1) for (c = 0; c < layer1_size; c++)
2     syn1neg[c + word*layer1_size] -= alpha * gsim[c];
```

Kontuan izan eguneraketa hori (A.9) ekuazioko ezkerraldeko batukaria dela, eskubiko eguneraketa jatorrizko A.1 kodeko 27. lerroan burutzen da, agerkidetzen begiztaren baitan.

[27] (A.9) ekuazioko eskubiko batukaria da.

[28] (A.10) ekuazioa.

[29] Behatutako hitz-testuinguru guztien eta azken horien lagin negatiboen kalkuluak bukatzen dituenean.

[30] Gogoratu A.1 kodean `word` indizeduna dela.