Konbinatu eta Irabazi! Hitzen Semantikaren Errepresentazio Osoagoaren Bila

Josu Goikoetxea Salutregi[1]

Eneko Agirre Bengoa[2] eta **Aitor Soroa Etxabe**[3]

[1] IXA taldea, UPV/EHU. josu.goikoetxea@ehu.eus
[2] IXA taldea, UPV/EHU. e.agirre@ehu.eus
[3] IXA taldea, UPV/EHU. a.soroa@ehu.eus

Laburpena

Artikulu honetan hitzen semantikaren errepresentazio osoagoa lortzeko algoritmo berria aurkezten dugu. Gure proposamenak bi metodo konbinatzen ditu: ezagutza-baseetan eta corpusetan oinarritutakoak. Hasieran, konbinatu barik, bi metodo horiekin antzekotasuna deituriko prozesu kognitiboa aztertu dugu, eta gero azken horiek uztartzen dituen algoritmoa enpirikoki frogatu dugu. Bi faseetan giza irizpideetan oinarritutako antzekotasun urre-patroiak hartu ditugu ebaluaziorako erreferentzia legez. Metodoak konbinatzerakoan baliabideak asko optimizatu ditugu, konbinatu barik lortutako emaitzak berdinduz edota gaindituz. Ondorioz, teknika horiek informazio semantiko osagarria dutela frogatu dugu.

Hitz gakoak: semantika, antzekotasuna, giza irizpideak, metodo konbinaketa

Abstract

In the article we present a novel algorithm which achieves a more integral representation of word semantics. Our proposal combines two methods: the ones based in knoledge-bases and in corpuses. In the beggining, we analize the cognitive process called similarity with those two methods, not combining them, and, afterwards, we empirically test the algorithm that yokes those methods. In both phases we have taken similarity gold-standards based on human criteria as a reference for evaluation. We have optimized the resources dramatically when combining methods, obtaining equal or better results than using them separatedly. Thus, we have proved those methods have complementary semantic information.

Keywords: semantics, similarity, human criteria, method combination

1 Sarrera eta motibazioa

Hitzen arteko antzekotasuna oso errotuta dago gizakiongan, guztiz naturala da gure egunerokotasunean, eta bere erabilera-esparrua oso zabala da; eguneroko hizkuntza arrunteko antzekotasun konkretuan hasi, eta literaturako eta olerkietako metaforen eta metonimien sofistikazio abstraktuetara arte joan daiteke. Eguneroko hizkuntzaren erabileraren adibide moduan, gizakiok badakigu *jaguar* eta *katu* hertsiki erlazionatuta daudela, eta, era berean, *jaguar* eta *dolar* hitzen artean harremanik ez dagoela. Ahalmen hori gakoa da hizkuntzan parte hartzen duten prozesu kognitiboen ulermenerako, eta Lengoaia Naturalaren Prozesamenduan (LNP) itzulpen automatikoan, esaterako, hainbat aurrerapen ahalbideratu ditu.

Egun, LNP arloan antzekotasunerako garatu diren algoritmo ahaltsuenak WordNet-en eta Wikipedian oinarritu dira (Agirre *et al.* (2010); Gabrilovich eta Markovitch (2007)), baina, azken aldian sare neuronalak gailentzen dabiltza (Mikolov *et al.* (2013); Collobert eta Weston (2008); Socher *et al.* (2011); Turian *et al.* (2010)). Antzekotasunerako ereduak ebaluatzeko hainbat aukera daude, baina denek ere giza irizpideak dituzte oinarri. Ohikoena hitzen antzekotasunari buruzko giza juzkuetan oinarritutako urre-patroiak erabiltzea da (Gabrilovich eta Markovitch (2007); Hill *et al.* (2014b)), eta hala egingo dugu lan honetan.

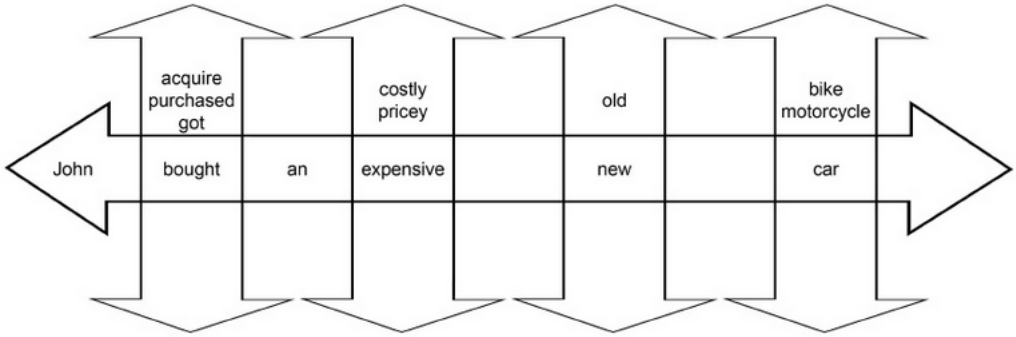
Ikerketa honen aztergai nagusia, baina, antzekotasunaren muina da, semantika. Azken horrek gizakiok

errealitatetik datorkigun informazioa egituratzeko hainbat prozesu kognitibotan parte hartzen du, eta, ondorioz, gure adimenaren gakoetako bat da.

Hala, goiko paragrafoan aipatutako metodoak semantika kudeatzeko bi estrategia dituzte:

- Sare neuronalek erlazio sintagmatikoak dituzte oinarri. Azken horiek, testuinguru berean sekuentzialki konbinatzen diren entitate linguistikoei aplikatzen zaizkie, eta elementuen posizionamendua da garrantzizkoena. 1 irudian ardatz horizontalaren konbinazio jakin baten entitateen arteko erlazioak izango lirateke; esaterako, *John bought an expensive new car* esaldiko hitzen artekoak.
- Ezagutza-baseek erlazio paradigmatikoak ustiatzen dituzte. Erlazio horiek testuinguru linguistiko berean ager daitezkeen entitateen artekoak dira eta 1 irudian ardatz bertikaleko hitzen artekoak dira. Erlazio bereko berben artean ordezkapenak egin daitezke; esaterako *John got an expensive new car* esaldia *John acquire an costly old bike* bihur daiteke.

1 Irudia: Ardatz sintagmatikoaren eta paradigmatikoaren adibidea.



Bi estrategia horiek nagusitu dira semantikaren errepresentazio abstraktuetan, eta gure ikerketan biak erabili ditugu; hasieran bakoitza bere aldetik, eta, gero, konbinatuta. Konbinaketa hori da, hain zuzen, gure ikerketaren ekarpenik garrantzitsuena.

Gauzak horrela, artikulu honetan gure esperimentazioaren bi fase deskribatzen dira. Lehenengoan, sare neuronaletan eta ezagutza-baseetan oinarritutako antzekotasun-esperimentuen emaitzak alderatuko ditugu, hiru hizkuntzatan. Bigarrenean, hitzen esanahiaren errepresentazioa hobetze aldera, aipatutako bi estrategiak konbinatuko ditugu hurrengo hipotesiari jarraiki: bi metodoen informazio semantikoa desberdina baina osagarria da, eta biak konbinatuta banaka baino emaitza hobeak lortuko dira. Azkenik, hipotesi hori esperimentuen bidez enpirikoki baieztatuko dugu.

2 Arloko egoera

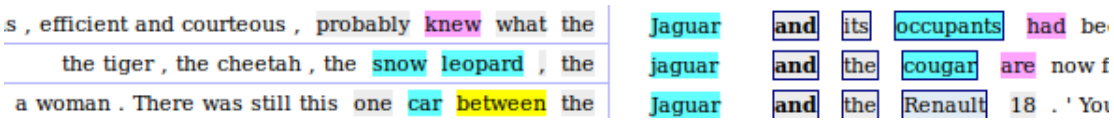
Atal honetan, lehenik, antzekotasuna aztertzeko bi eredu azaltzen ditugu; hots, ezagutza-baseetan oinarritutakoak eta corpusetan oinarritutakoak. Gero, eredu horiek ebaluatzeko hitz-pareen urre-patroiak izango ditugu hizpide. Azkenik, antzekotasuna neurtzeko algoritmoa azalduko dugu.

2.1 Corpusetan oinarritutako teknikak

Azken urteotan sare neuronalak hitzen semantikarekin lan egiteko tresna baliagarria bilakatu dira. Garuneko neuronen eredu biologikoa jarraitzen dute, eta egun garatutako guztiek bi ezaugarri partekatzen dituzte; alde batetik, corpus ez etiketatuetatik erauzten dute hitzen esanahia, eta, bestetik, semantika modu distribuzionalean kudeatzen dute. Bada, sare neuronalek Hipotesi Distribuzionala (Harris, 1954) jarraitzen dute, eta azken horrek zera dio: *esanahi antzeko hitzek testuinguru berean agertzeko joera dute.* Esaterako, 2 irudian *jaguar* hitzaren bi testuinguru posible agertzen dira; hots, animaliaren adieraren ingurukoa, eta autoaren ingurukoa.

[1] Jatorria: http://corpus.byu.edu/bnc/

2 Irudia: *Jaguar* hitza testuinguru desberdinetan.[1]



Sare neuronalen ereduei dagokienez, hainbat aztertu ditugu ((Collobert eta Weston (2008)); (Turian *et al.* (2010)); (Socher *et al.* (2011));(Mikolov *et al.* (2013))), baina, emaitzarik onenak Mikolov-en ereduarekin[2] lortu dira, eta, hortaz, azken hori erabili dugu. Mikolov-en eredua, corpus erraldoietatik abiatuta (10^9 berbatik gora), hitzen kalitatezko esanahiak modu arinean ikasteko pentsatuta dago, kostu konputazional baxuarekin eta zehaztasun handiarekin. Eredu horretan bi proposamen daude: Continuous Bag-of-Words (CBOW) eta Skip-gram.

Hipotesi Distribuzionalari lotuta, CBOW ereduan entrenamenduko irizpidea hurrengo da: aurreko eta ondorengo hitzak jakinik, erdikoa zuzen iragarri. Esaterako, 1 adibideko esaldia emanik, CBOW ereduak *the scared* eta *jumped inside* hitzen bektore-adierazpenak dakizki, eta erdiko hitzarena lortu behar du. Operazio hori corpuseko berba guztiekin egiten da. Skip-gram ereduak, ordea, erdiko hitza emanik aurrekoak eta ondorengoak aurresaten ditu. Hala, aurreko adibideari jarraiki, *jaguar* hitzetik *the scared* eta *jumped inside* aurresan beharko ditu.

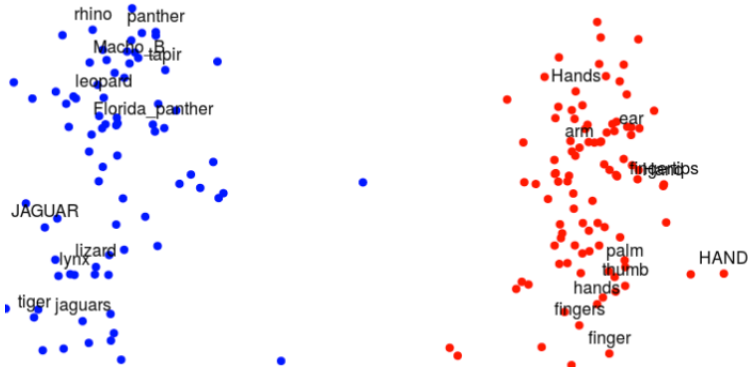
(1) *the scared* **jaguar** *jumped inside*

Gauzak horrela, sare neuronalak bektore-adierazpen deituriko errepresentazioak ikasten ditu; hots, balio eskalarrez osatutako bektore oso trinkoak. Bektore-adierazpenaren dimentsio bakoitzak esanahiaren ezaugarri semantiko edota sintaktiko bat adierazten du, modu abstraktuan. Modu intuitiboagoan ulertzeko, bektoreen balio eskalarrak N dimentsiotako espazio bateko koordenatu legez har daitezke, eta esanahi bakoitza posizio jakin batean ezartzen dute. Esaterako, gure ikerketaren aurreneko fasean *jaguar* hitzaren bektorearen lehenengo 8 dimentsioek hurrengo itxura daukate :

(2) *jaguar = [0.030852 0.049318 -0.035806 0.029951 0.052470 -0.029501 0.026686 0.005883 ...]*

Osagai Nagusizko Analisiaren (ONA) bidez N dimentsioko bektoreak bitara pasa daitezkeenez, 3 irudian *jaguar* eta *hand* bektoreetatik auzokide semantikoak (esanahi antzekodunak) irudikatu ditugu; hots, hitz horien erlazio paradigmatikoetan oinarritutako auzokide semantikoak irudikatu ditugu. Irudi horretan Hipotesi Distribuzionalean esandakoa grafikoki ikus daiteke, antzeko testuinguruetatik erauzitako hitzak multzokatuta agertzen baitira kolore bereko puntuetan.

3 Irudia: *Jaguar* eta *hand* hitzen auzokide semantikoak Skip-gram-ekin erauziak, urdinez eta gorriz, hurrenez hurren.



[2]https://code.google.com/p/word2vec/

2.2 Ezagutza-baseetan oinarritutako teknikak

Azken urteotan grafoetan oinarritutako teknikek oihartzun handia izan dute LNP komunitatean, grafoen ezaugarri estrukturalak bilatu eta ustiatzen dituztelako. Gainera, teknika horiek grafoa bere osotasunean aztertzen dutenez, soluzio globalak eta optimoak lortzeko gai dira.

Ikerketa honetan eginiko esperimentuetan bi ezagutza-base, WordNet[3] eta Wikipedia, erabili ditugu. Lehenengoa psikolinguistikako teoriak frogatzeko asmoarekin sortu zen, eta ingelesaren datu-base lexikala da; hitzak synset deituriko sinonimo multzoetan elkartzen ditu, eta azken horietako bakoitzak kontzeptu desberdin bat adierazten du. Gainera, sinonimo-multzo horiek sare bat osatzen dute, erlazio semantikoez eta lexikalez lotutakoa.

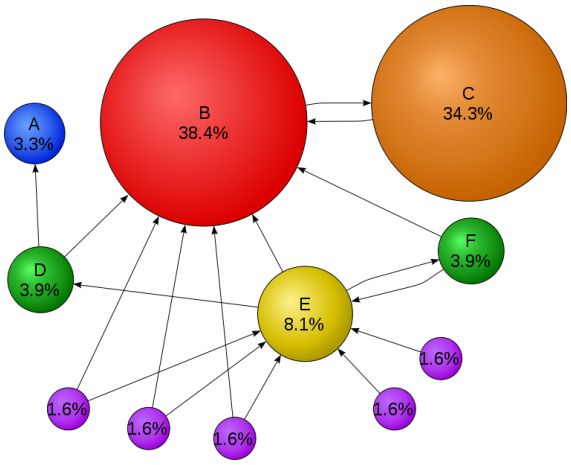
Wikipediari dagokionez, milaka kolaboratzailek sortutako online entziklopedia erraldoia eta eleanitza da, hainbat gairen inguruko artikuluak dituena. Wikipedia-n, artikuluez gain, artikuluen artean hainbat erlazio mota daude.

Gauzak horrela, gure ikerketetan UKB[4] programa-bilduma erabili dugu. UKB-k ezagutza-baseak grafo bezala ulertuko ditu; hau da, WordNet-eko eta Wikipedia-ko kontzeptuak grafoko adabegi legez hartzen ditu, eta azken horien arteko erlazioak adabegien arteko ertz moduan.

UKB-k sarrera-hitz bat jasotzerakoan, Page Rank Pertsonalizatua (PRP) (Agirre eta Soroa, 2009) deituriko algoritmoa aplikatzen dio. Hasteko, sarrerari dagokion adabegitik ausazko ibilbide bat abiatzen du, betiere erlazionatuta dauden adabegien artetik pasatuz. PRP-k adabegi bakoitzari bere ertz kopuruaren araberako probabilitate bat esleitzen dio, eta ibilbidea pisu handiagoa dutenetara bideratu.

Ausazko ibilbideak konbergitzean, sarrerako hitz bakoitzari PRP-k probabilitateez osaturiko bektore (PPB) bat esleitzen dio, eta bektore hori da, hain zuzen, hitzaren esanahia adieraziko duena. 4 irudiak grafikoki adierazten du grafo sinple batean PPB pisuak zeintzuk liratekeen.

4 Irudia: PageRank pisuak grafo sinple batean.[5]



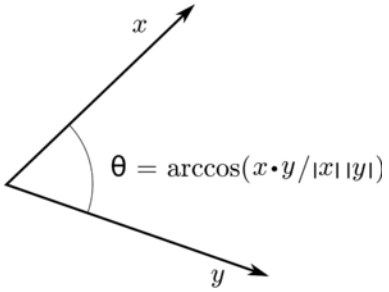
Azpimarratzekoa da PPB-ek bektore-adierazpenek baino askoz ere dimentsio gehiago dituztela, UKB-k grafoko adabegi bakoitzeko dimentsio bana esleitzen baitu. PPB-en tamaina desabantaila da, baina, beranduago ikusiko dugunez, gure esperimentuen azkenengo fasean irtenbide bat emango diogu.

2.3 Hitzen arteko antzekotasuna

Antzekotasuna kalkulatzeko hainbat teknika daude, baina, bektoreekin lan egitean kosinuaren formula estandarra da. Algoritmo hori aljebra linealeko biderketa eskalarrean oinarritzen da, eta, beraz, eragiketaren emaitza eskalar bat da; bi bektoreren arteko angeluaren kosinua, hain zuzen.

[3] http://wordnet.princeton.edu/
[4] `http://ixa2.si.ehu.es/ukb/`
[5] Jatorria: http://en.wikipedia.org/wiki/PageRank

5 Irudia: Bi bektoreren arteko cosinua.



Behin esanahi-bektoreak esku artean (2.1 ataleko bektore-adierazpenak edota 2.2 ataleko PPB-ak), bi hitzen arteko antzekotasuna bektore horien arteko kosinuarekin neur daiteke; zenbat eta angelu txikiagoa izan, orduan eta antz gehiago izango dute hitzek. Bestela esanda, eragiketaren emaitza zenbat eta 1etik hurbilago egon, orduan eta antzekotasun semantiko handiagoa. Hala, 1 ekuazioan x eta y hitzen esanahien distribuzioa x_i eta y_i pisuko $\vec{x}$ eta $\vec{y}$ bektoreak bezala hartzen dira, hurrenez hurren.

$$antzekotasuna(\vec{x},\vec{y}) = \cos(\theta((\vec{x},\vec{y}))) = \frac{\vec{x}.\vec{y}}{\|\vec{x}\|\|\vec{y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

Demagun *jaguar* eta *cat* hitzak erauzi ditugula Mikolov-en sare neuronalarekin, 300 dimentsiorekin. Bi kontzeptu horien antzekotasuna 3 adibidean legez kalkulatuko litzateke:

$$(3)\ antzekotasuna(\vec{jaguar},\vec{cat}) = \frac{\sum_{i=1}^{300} jaguar_i cat_i}{\sqrt{\sum_{i=1}^{300} jaguar_i^2}\sqrt{\sum_{i=1}^{3} cat_i^2}} = 0.41$$

Esan bezala, 3 adibideko emaitza zerotik baterako tartean ulertu behar dugu; Mikolov-en ereduaren aburuz, beraz, hitz horien artean badago antzekotasun esanguratsua.

2.4 Hitz-pareen urre-patroiak

WordSim353 eta SimLex999 datu-multzoak antzekotasuna neurtzen duten algoritmoak ebaluatzeko erabiltzen dira. Bi datu-multzo horiek giza irizpideekin osatuta daude, eta hainbat ikerketek giza irizpideek bata bestearekin korrelazio altua dutela adierazi dute. Horrexegatik, hain zuzen, WordSim353 eta SimLex999 ebaluaziorako baliabide egokiak dira.

Datu-multzo horiek eskuz eginikoak dira, eta ingelesezko hitz-bikote bildumez osatuta daude, gizakiek pare bakoitzaren inguruan emandako antzekotasun- edota ahaidetasun-irizpideen balio numerikoekin: zerok antzekotasunik edo ahaidetasunik eza adieratzen du, eta hamarrerek hitz bera edo sinonimoak direla. WordSim353 datu-multzoa 353 hitz-bikotez osatuta dago eta SimLex999 999rekin.

6 Irudia: WordSim353 datu-multzoko 5 bikote.

```
cup     substance  1.92
cup     liquid     5.90
jaguar  cat        7.42
jaguar  car        7.27
energy  secretary  1.81
```

Behin hona helduta, garrantzitsua da antzekotasuna eta ahaidetasuna terminoak desberdintzea. Antzekotasunak hitzen arteko sinonimia (hitz/berba) eta hiperonimia/hiponimia (taxi/auto) hartzen ditu bere baitan, eta ahaidetasunak, sinonimia eta hiperonimia/hiponimiaz gain, meronimia (esku/hatz), antonimia (hotz/bero) eta asoziazioa. Bada, WordSim353 urre-patroian, antzekotasuna eta ahaidetasuna nahastuta agertzen dira; SimLex999-n, ordea, antzekotasuna soilik agertzen da. Artikulu honetan antzekotasuna terminoa bakarrik erabili dugula azpimarratu behar da, biak berdin prozesatzen baitira. Hala ere, etorkizunean bi terminoak bereizteko eta neurketak desberdintzeko asmoa daukagu.

3 Esperimentuak

3.1 Esperimentu eleanitza: metodoak konbinatu barik

Esperimentuen aurreneko fasean antzekotasunaren ebaluazio eleanitza egin dugu, eta 2.4 puntuan azaldutako WordSim353 datu-multzoa erabili dugu ebaluaziorako urre-patroi bezala. Eleaniztasun hori dela-eta, hiru datu-multzorekin egin dugu lan, euskarazko eta gaztelerazko urre-patroiak ingelesekoaren itzulpenak izanik. Arrazoi beragatik, hiru hizkuntzen bektore-adierazpenak erauzteko hiru corpus erabili ditugu: euskarazko corpusa Elhuyar fundazioak utzitako corpusa da, artikulu zientifikoz osatua dago eta $1.5 \cdot 10^8$ berbaz osatua; ingeleseko bektore-adierazpenak 10^{11} berbako Google News corpusetik erauzita daude, eta Google Code-tik[6] jaitsi ditugu zuzenean; gaztelerazkoa QTLeap protiektuan erabilitako corpusen bilduma da, $0.8 \cdot 10^9$ hitz ingurukoa.

Antzekotasuna bektore-adierazpenekin, WordNet-en eta Wikipedia-ren PPB-ekin kalkulatu dugu, eta hiru baliabide horiek euskaraz, gazteleraz eta ingelesez ebaluatu ditugu. WordNet 3.0g[7] bertsioa erabili dugu. Grafo modua hartuta, bertsio horretan 117.522 adabegi (synset) eta 525.356 ertz semantiko daude. Hiru hizkuntzarekin ibili bagara ere, hiztegiak aldatzen dira soilik; grafoa berbera da euskaraz, ingelesez eta gazteleraz. Arrazoi beragatik, Wikipedian hiru iraulketa erabili ditugu: euskarazkoa $3 \cdot 10^6$ adabegirekin eta $160 \cdot 10^3$ ertzekin, ingelesezkoa $16 \cdot 10^6$ adabegirekin eta $3 \cdot 10^6$ ertzekin eta gaztelerazkoa $15 \cdot 10^6$ adabegirekin eta $450 \cdot 10^3$ ertzekin. Ebaluazioak egiteko, antzekotasun-neurketen eta hizkuntza guztien WordSim353 urre-patroien arteko korrelazioa kalkulatu dugu Spearman[8] bidez.

1 Taula: WordSim353 eleanitzeko Spearman balioak

	Euskara	Gaztelera	Ingelesa
Bektore-adierazpenak	0.3329	0.3598	0.686
PPB Wordet	0.3712	0.3929	0.683
PPB Wikipedia	**0.4212**	**0.4645**	**0.7274**

1. taulako emaitzei so, argi ikusten da ingelesak korrelaziorik altuenak dauzkala hiru baliabideetan eta Wikipedia gailentzen dela hiru eleetan. Emaitzen desoreka horren arrazoiak hurrengoak dira: ingeleseko corpusak kalitate hobea eta tamaina handiagoa dauka, eta grafoak handiagoak dira; urre-patroien itzulpenetan ingelesaren konnotazio kultural desberdinek eragina dute beste bi hizkuntzen ebaluazio-emaitzetan. Ikerketako fase horren helburu nagusia sare neuronalek eta ezagutza-baseek gizakion antzekotasun-irizpideekin duten korrelazioak neurtzea eta erkatzea izan da, hiru hizkuntzatan.

3.2 Esperimentu elebakarra: metodoak konbinatuta

Orain arte sare-neuronalak eta ezagutza-baseak bakoitza bere aldetik erabili ditugu. Hala ere, bi paradigma horietatik erauzitako informazioa osagarria dela uste dugu, eta, gure aburuz, bi bideak konbinatzea lortuz gero banaka lortutako emaitzak hobetu beharko genituzke. Hipotesi horri jarraiki, esperimentuen hurrengo fasean hitzen semantikaren errepresentazioa hobetzea izan dugu helburu, eta, hori lortze aldera, aipatutako bi estrategiak uztartu ditugu.

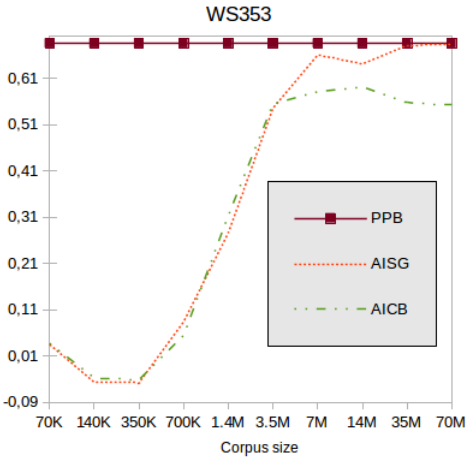
Horretarako, WordNet-eko hitzen gainean ausazko ibilbideak egin ditugu, eta ibilbide horietan igarotako hitzekin testu corpus bat eratu dugu; gure ustez, corpus horrek inplizituki WordNet-eko erlazio paradigmatikoen informazioa gordetzen du. Gauzak horrela, testu corpus hori Mikolov-en eredua elikatzeko erabili dugu, eta, ondorioz, erlazio paradigmatiko inplizituak sintagmatikoak balira bezala prozesatu ditugu. 1 irudiko bi ardatzak, nolabait, batu egin ditugu. Lortutako bektore-adierazpenak PPB-ak baino askoz trinkoagoak dira (300 dimentsio vs. 10 milako batzuk), eta, ikusiko dugunez, oso emaitza onak emango ditu antzekotasun atazetan.

Esperimentu honetan CBOW eta Skip-gram erabili ditugu, parametro berdinekin, eta corpusaren tamaina handitzen joan gara $70 \cdot 10^3$-tik $70 \cdot 10^6$ lerrora arte. Corpusa handitu ahala Spearman balioak

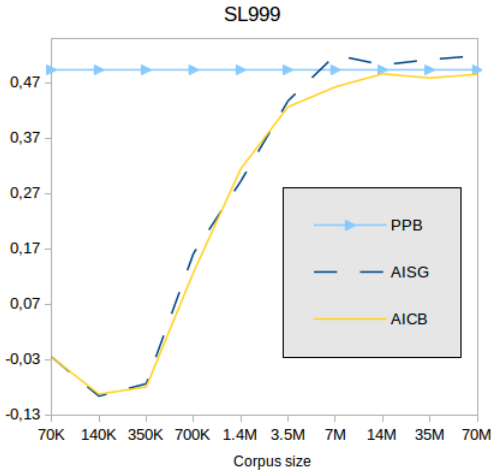
[6] https://docs.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit
[7] http://wordnet.princeton.edu/glosstag.shtml
[8] http://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

kalkulatu ditugu, eta gure metodoaren ikasketa-kurba eratu dugu. Aurreko fasean bezalako metodologia jarraitu dugu, baina ingelesarekin soilik eta bi datu-multzorekin, SimLex999-rekin eta WordSim353-rekin.

7 Irudia: Ikasketa-kurba WS353-rekin.



8 Irudia: Ikasketa-kurba SL999-rekin.



7 eta 8 irudiek WordNet gainean ausazko ibilbideekin eratutako corpusei CBOW (AICB) eta Skip-gram (AISG) aplikatuta lortutako Spearman balioen bilakaera erakusten dute, SL999 eta WS353 datu-multzoekin. Ikasketa-kurbek $7 \cdot 10^6$ testuingurutan konbergitzen dute, PPB-en besteko emaitza lortuz WordSim353-rekin, eta SL999-koak gaindituz. Kobergentzia horretako balioak 2 taulan agertzen dira. Gauzak horrela, gure metodoari jarraiki, sare neuronalek WordNeteko informazio semantikoa kodetzeko gai direla frogatzen dugu.

2 Taula: AISG-rekin eta AICB-rekin lortutako Spearman korrelazioak ingelesez, PPB eta Skip-gram soilekin alderatuta.

	SL999	WS353
Bektore-adierazpenak	0.442	**0.686**
AISG	**0.520**	0.683
AICB	0.486	0.591
PPB WordNet	0.493	0.683

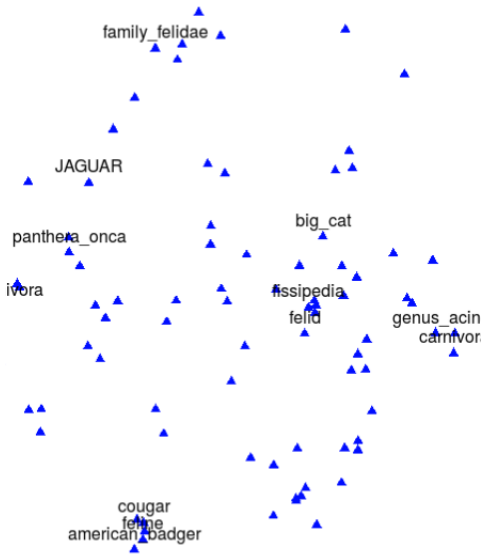
9 eta 10 irudietan *jaguar* bektorearen hitz antzekoenak irudikatu dira ONA erabilita; alde batetik, *word2vec* tresnarekin erauzitako bektoreak (auzikude sintagmatikoak); bestetik, AISG bidez erauzitakoak. 9 irudian corpus natural batetik ateratako hitzak agertzen dira, *jaguar* hitzarekin erlazio sintagmatikoak dituztenak: hala nola, *coyote*, *rhino* edota *owl* modukoak. 10 irudikoak, ordea, WordNet-teko elementuak dira, *jaguar*-ekin erlazio paradigmatikoak dituztenak: *big_cat* eta *family_felidae* dira adibiderik argienak. Bada, AISG eta AICB metodoen bidez, WordNet-eko hitzak (inplizitoki erlazio paradikmatikoak dituztenak) erlazio sintagmatikoak balituzte bezala prozesatu ditugu; hortik gure metodoaren ekarpen nagusia.

SimLex999-aren artearen egoera (Hill *et al.*, 2014a) berdindu egin dugu, egun 0.52 baita. WordSim353-en (Radinsky *et al.*, 2011), ordea, ez gara 0.8 baliora heldu. Hala ere, kontuan izan behar dugu esperimentu hauek paradigma osagarriak uztartzeko aurreneko urratsak baino ez direla. Bide honetatik, etorkizunean, konbinazio sofistikatuagoekin artearen egoera hobetu dezakegu.

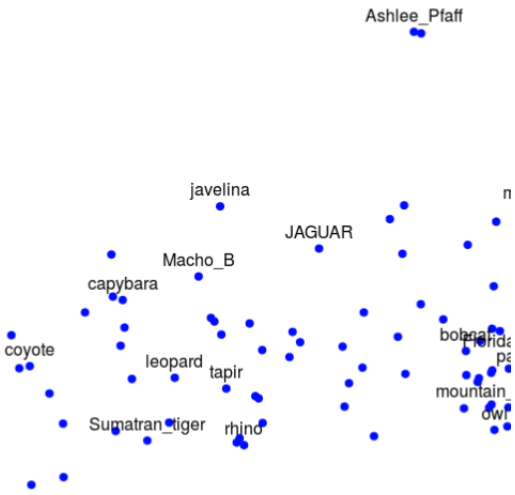
4 Ondorioak eta etorkizuneko egitekoak

Gure ikerketetan antzekotasuna modelatzeko hainbat teknika erabili ditugu, eta guztiak giza irizpidean oinarrituta ebaluatu. Egun egindako lanekin erkatuta, gure esperimentuen lehenengo fasean ezagutza-baseak eta bektore-adierazpenak baliabide bezala erabili izana, hiru hizkuntzatan, kuantitatiboki balio-

9 Irudia: *Jaguar* hitzaren auzokideak, AISG bidez erauzita



10 Irudia: *Jaguar* hitzaren auzokide sintagmatikoak, bektore-adierazpenen bidez erauzita



tsua izan da. Hala ere, gure ikerketa-lerroaren ekarpen nagusia bigarren fasean dago, bi paradigma horien konbinaketa oso gutxi landutako alorra baita (Weston *et al.* (2013); Wang *et al.* (2014)).

Aurreneko faseko emaitzek hizkuntzen baliabideen arteko desoreka agerian utzi dute, Wikipediaren eta ingelesaren faboretan. Urre-patroien itzulpena dela-eta, konnotazio kulturalen garrantziaz jabetu gara, ebaluazioko emaitzetan zarata-iturri izan daitezkeela argi ikusi dugulako. Esan gabe doa informazio hori eleaniztasunak eta eleen gurutzaketak ahalbidetu digula.

Bigarren fasean ezagutza-baseen egitura eta sare neuronalak uztartzen dituen metodo berria aurkeztu dugu, bektore-adierazpen konbinatua sortzen duena. Gure antzekotasun ebaluazioetan PPB-ekin lortutakoen emaitzak berdindu ditugu, baina milaka dimentsioko bektoreak barik askoz trinkoagoak (300 dimentsiokoak) erabilita. Are gehiago, hitzen errepresentazio horien informazioa sare neuronaletan eta ezagutza-baseetan oinarritutakoekin osagarria da, eta, ondorioz, azken horiek banaka erabilita baino emaitza hobeak lortzen ditugu. Beraz, gure metodo berri honek ateak irekitzen dizkio orain arte banatuta egon diren bi estrategia horien konbinaketei.

Egiteke dauden esperimentuei begira, hainbat puntu geratzen dira irekita. Hasteko, euskarazko eta gaztelerazko baliabideak hobetzea beharrezkoa iruditzen zaigu, eta hizkuntza-gurutzaketaren potentziala ustiatzearen garrantziaz jabetu gara. Gure ustez, estrategien konbinaketaren eta eleaniztasunaren nahasketak aberastasuna eta sakontasuna emango dio ikerketari. Hori erdieste aldera, nahitaezkoa da euskarazko eta gaztelerazko urre-patroiak osatzeko datuak biltzea, eta kalitate eta tamaina hobeko corpusak lortzea.

Paradigmen konbinaketei begira, hurrengo esperimentuetan ausazko ibilbideekin Wikipedian oinarritutako corpusa osatu nahi dugu, eta konbinaketen filosofia beste esparru batzuetara hedatu; alde batetik, iturri desberdinetatik lortutako bektore-adierazpenak konbinatu nahi ditugu, eta, bestetik, corpusak. Esaterako, corpus naturalak eta WordNet gainean AISG-rekin lortutako corpusak nahastu. Horrez gain, *word2vec* tresnaren parametroak optimizatuta bektore-adierazpenen kalitatea hobetzea espero dugu.

Amaitzeko, esan beharra dago ikerketa hau hitzen semantikaren errepresentazio osoagoa erdiesteko ahalegina dela. Esanahi osoagoaren bilaketa horretan, aipatutako bi paradigmen azaleko desberdintasunak alde batera utzi ditugu eta semantika ulertzeko bi ikuspuntuen ustezko osagarritasunean oinarritu gara. Bestela esanda, gure ikerketan erlazio semantikoek eta paradigmatikoek batak bestearen gabezia semantikoak betetzen dituela izan dugu hipotesi. Orain arteko esperimentuen emaitzei so badirudi norabide zuzenetik goazela, eta semantika ulertzeko metodo desberdinen konbinaketak hainbat aukera ireki ditzakeela.

Erreferentziak

Agirre, Eneko, Montse Cuadros, German Rigau, eta Aitor Soroa. 2010. Exploring Knowledge Bases for Similarity. In *LREC*.

——, eta Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 33–41. Association for Computational Linguistics.

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

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

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

Hill, Felix, KyungHyun Cho, Sébastien Jean, Coline Devin, eta Yoshua Bengio. 2014a. Not all neural embeddings are born equal. *CoRR* abs/1410.0718.

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

Mikolov, Tomas, Wen-tau Yih, eta Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, 746–751.

Radinsky, Kira, Eugene Agichtein, Evgeniy Gabrilovich, eta Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, WWW '11, 337–346, New York, NY, USA. ACM.

Socher, Richard, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, eta Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 151–161. Association for Computational Linguistics.

Turian, Joseph, Lev Ratinov, eta Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 384–394. Association for Computational Linguistics.

Wang, Zhen, Jianwen Zhang, Jianlin Feng, eta Zheng Chen. 2014. Knowledge graph and text jointly embedding. 1591–1601.

Weston, Jason, Antoine Bordes, Oksana Yakhnenko, eta Nicolas Usunier. 2013. Connecting language and knowledge bases with embedding models for relation extraction. *arXiv preprint arXiv:1307.7973* .

5 Eskerrak

Eskerrak eman nahi dizkiot EHU-ko IXA taldeari ikerketa hau aurrera ateratzeko baliabideak eta azpiegitura eskaini dizkidalako.

Eskerrak baita ere Iraide Zipitriari, bere euskarazko eta gaztelerazko datu-multzoek proiektuaren aurreneko fasean eleaniztasuna jorratzeko aukera eman baitidate.