Hedapen Semantikoa Informazioaren Berreskurapenean

Arantxa Otegi, Eneko Agirre, Xabier Arregi*

IXA Taldea, Informatika Fakultatea
(UPV/EHU)

* arantza.otegi@ehu.es

Jasoa: 2014-06-13 Onartua: 2014-09-22

Laburpena: Informazioaren berreskurapena (IB) erabiltzaile baten informazio-beharra aseko duten dokumentuak bilatzean datza. Beraz, IB sistemak erabiltzaileari laguntza emango dio dokumentu adierazgarriak, alegia, erabiltzaileak behar duen informazioa eduki dezaketen dokumentuak, topatzeko. Horretarako, erabiltzaileak egindako kontsultan oinarritzen gara. Kontsultaren eta dokumentuen arteko parekatze arazoa deiturikoa da IB sistemek aurre egin behar dioten arazo nagusienetako bat: dokumentu bat kontsulta baterako adierazgarria izan daiteke, nahiz eta bietan erabilitako hitzak guztiz berdinak ez izan, eta, alderantziz, dokumentu bat ezadierazgarria izan daiteke kontsulta baterako, nahiz eta termino komun batzuk eduki. Arazo hauek hitzen sinonimiaren eta anbiguotasunaren kausaz gertatzen dira. Lan honetan, kontsulten eta dokumentuen hedapenak egin eta aurre egingo diogu parekatze arazoari, hizkuntzaren prozesamenduko hitzen adiera-desanbiguazioa eta ahaidetasun semantikoa erabiliz. Hiru datu multzotan egindako esperimentu eta analisiek erakusten dute proposatutako hedapen-metodoek parekatze arazoari aurre egiteko balio dutela eta, ondorioz, baita IB sistemaren eraginkortasuna hobetzeko ere.

Hitz gakoak: informazioaren berreskurapena, hedapen semantikoa.

Abstract: *Information retrieval* (IR) aims at searching documents which satisfy the information need of an user. In that way, an IR system informs the user about relevant documents, that is those documents that contain the information they need as formulated in the query. One of the main problems is the so-called vocabulary mismatch problem between query and documents: some documents might be relevant to the query even if the specific terms used differ substantially, or some documents might not be relevant to the query even if they have some terms in common. The former is because several words or phrases can be used to express the same idea or item (synonymy). The latter is caused by ambiguity, where one word can have more than one interpretation depending on the context. In this work, we expand queries and documents making use of two NLP techniques, word sense disambiguation and semantic related-

ness. Our extensive experiments on three datasets show that the expansion methods explored in this dissertation help overcome the mismatch problem, consequently improving the effectiveness of an IR system.

Keywords: information retrieval, semantic expansion.

1. **SARRERA**

Informazio-beharrak asetzeko maiz jotzen dugu dokumentu elektronikoen biltegietara gaur egun. Joera hau oso ohikoa da mundu zabaleko dokumentuak Internet sareak esku-eskura jarri dizkigunetik. Hain zuzen ere, informazioa dokumentu-biltegietan bilatzeko zeregin horri *Information Retrieval* deitzen zaio.

Termino hori 1950ean erabili zen lehen aldiz [1], eta geroztik erabat *errotu zen*. Euskaraz Informazioaren Berreskurapena (aurrerantzean IBa) deritzo arlo honi.

Horrela, bada, IB sistema bat dokumentuetako informazioa biltegiratu eta kudeatzen duen software-programa da [2]. Behar duen informazioa topatzen lagunduko dio erabiltzaileari sistemak, informazio hori eduki dezaketen dokumentuen berri emanez. Kontuan izan behar da horrelako sistemek ez dutela informazioa esplizituki itzultzen edo erabiltzailearen galderari zuzenean erantzuten, eta dokumentuak berreskuratu edo iradoki besterik ez dute egiten.

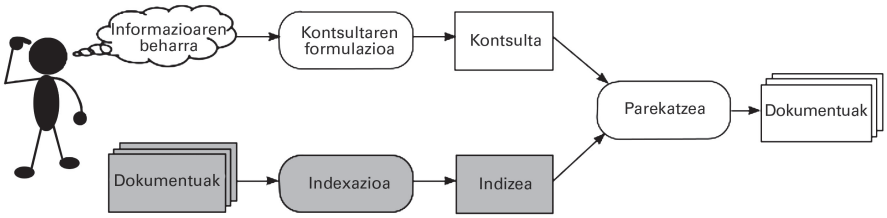
1.1. **IB sistemen nondik norakoak**

IB sistema batek hiru prozesu nagusi egiten ditu (ikus 1. irudia):

(i) Indexazioa: dokumentuen irudikapena gauzatzen da, indizea(k) sortuz. Indizea bilaketa azkarrak egiteko aukera ematen duen datu-egitura da. Indexazioa lineaz kanpo (*offline*) egikaritzen da, eta dokumentu bilduma aldatzen ez bada behintzat, behin egitea nahikoa da.
(ii) Kontsultaren formulazioa: erabiltzailearen informazio-beharra kontsulta baten bidez adierazten da.
(iii) Parekatzea: kontsulta dokumentuen irudikapenarekin parekatzen da, hau da, indizearekin. Parekatzea lortzen dugunean, pentsa dezakegu kontsultako terminoek eta indizeko terminoek gertutasun handia dutela elkarrekin. Hala, parekatzearen emaitza gisa, dokumentu arrakastatsuen multzoa aukeratzen da, dokumentu bilduma osoaren azpimultzoa izango dena.

Irteerako azpimultzo horretako dokumentu batzuek, ziur aski, erabiltzailearen informazio-beharrari erantzungo diote; dokumentu horiei do-

kumentu adierazgarri deitzen zaie. IB sistema perfektu batek dokumentu adierazgarriak baino ez lituzke berreskuratu beharko, ezadierazgarriak baztertuz. Alabaina, ez dago sistema perfekturik (geroago ikusiko dugu zein diren sistema hauen gabezietako batzuk). Gaur egungo sistemetan dokumentu-zerrenda ordenatu bat itzultzea da ohikoena; horrela, zerrendaren hasieran jartzen dira ustez erabiltzaileari gehien interesatuko zaizkion dokumentuak, alegia, sistemaren iritziz adierazgarrienak direnak.



1. irudia. IB sistema baten prozesua modu eskematikoan. Grisez markatuta dagoena lineaz kanpo gauzatzen da.

Prozesu horiek konputagailuen bidez guztiz automatikoki egikaritzeko ideia Bushek [3] proposatu zuen lehen aldiz 1945ean. Geroztik, ideia abstraktu izatetik–hainbat esparrutan erabiltzera pasa da. Esate baterako, IB sistemen adibide garbiak dira hain ezagunak eta erabiliak diren Google[1] eta Yahoo![2] bezalako web-bilatzaileak.

IB sistemek zabalkunde handia lortu dute, eta, jakina, arlo horretako ikerketa-gaiek ere indarra hartu dute. Testu-dokumentuen berreskurapenari dagokionez, honako ikergai hauek nabarmendu dira azkeneko urte hauetan: adierazgarritasuna mailakatzeko funtzioen (ranking-funtzioen) eraginkortasuna, sistemaren errendimendua (erantzun-denbora, indexatzeko denbora...), dokumentu edo datu berriak indizean txertatzeko azkartasuna, sistemaren eskalagarritasuna (datu edo erabiltzaile kopuruarekiko), aplikazio berrietara egokitzeko gaitasuna, ebaluazioa (metrikak eta ebaluazio-metodologiak) eta azkenik, kontsulten eta dokumentuen parekatze arazoa.

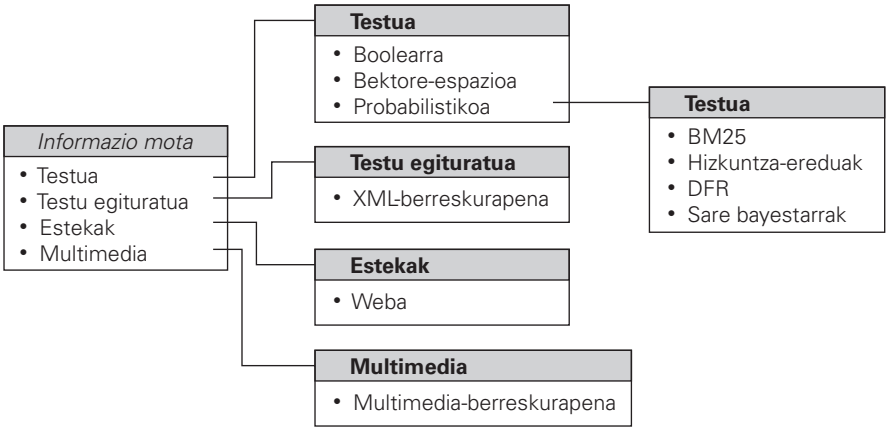
Hain zuzen ere, azken horretan, hau da, parekatze arazoan sakonduko dugu hemendik aurrerako ataletan.

[1] http://www.google.es/
[2] http://es.yahoo.com/

2. **BERRESKURAPEN-EREDUEN BILAKAERA**

IB sistema baten ezaugarri behinena berreskurapen-eredua da. Erabiltzaileak egindako kontsulta bat emanik, berreskurapen-eredu baten arabera erabakitzen da zein diren dokumentu adierazgarriak erabiltzailearentzat, horretarako arrazoizko biltegiratze-espazioa erabiliz eta erantzuna arrazoizko denboran emanez.

2. irudian ikus daiteke IB ereduen sailkapen bat, berreskuratzen den dokumentu motaren arabera antolatua: testuak, hipertestuak (estekak) eta multimedia. Testua berreskuratzeko ereduetara mugatuko gara atal honetan, eta haien artean hiru familia nagusiak aurkeztuko ditugu: eredu boolearrak, bektore-espazioaren ereduak eta probabilitate-ereduak. [5] liburuan aurki daitezke eredu nagusienen eta jorratuko ez ditugun beste hainbat ereduren zehaztapen gehiago.



2. irudia. IB ereduen sailkapena.

2.1. **Eredu boolearrak**

Boolearra izan zen lehenengo bilatzaileek erabiltzen zuten berreskurapen-eredua. Multzo-teorian eta logika boolearrean oinarritzen da eredu hau. Bai kontsultak bai dokumentuak terminoen multzotzat hartzen dira. Logika boolearreko *AND*, *OR* edo *NOT* eragileekin hainbat termino elkartuz osatzen dira kontsultak. Kontsultako adierazpen logikoa egiazko egiten duten dokumentuak izango dira berreskuratuko direnak (adierazgarriak), eta gainontzeko dokumentuak baztertu egingo dira (ezadierazgarriak). Hortaz, eredu honetan dokumentuak bi multzotan sailkatzen dira: adierazgarriak eta ezadierazgarriak. Beraz, ez da dokumentuen mailakatzerik edo rankingik egiten.

Gaur egun oraindik sistema batzuek eredu hau erabiltzen badute ere, 60ko hamarkadatik aurrera indar handia hartu zuten bestelako eredu batzuek. Eredu estatistikoak dira ondoren ikusiko ditugunak, dokumentuetako terminoen agerpen kopuruak hartzen baitituzte kontuan rankingak egiteko. Kontsulta-adierazpenak automatikoki sortzen dira, hots, erabiltzaileak informazio-beharra hizkuntza arruntean idatzi ahal izango du, edo termino gutxi batzuk idatzi ahal izango ditu, eragile logikoak erabili beharrik izan gabe.

2.2. **Bektore-espazioaren ereduak**

Luhn izan zen informazio-bilaketaren prozesuari ikuspuntu estatistiko bat eman behar zitzaiola proposatzen lehena 1957an [6]. Aurrerago, antzekotasun-irizpideari eutsiz, Saltonek [7] eredu sendoago bat garatu zuen: bektore-espazioaren eredua (ingelesez *vector space model* moduan ezagutzen dena). Eredu honetan kontsultak eta dokumentuak *N* dimentsioko bektore moduan adierazten dira, *N* bildumako termino kopurua izanik. Alegia, termino bakoitzeko bektoreko elementu bat izango dugu, eta elementu bakoitzak pisu bat izango du.

Behin kontsultari eta dokumentu bakoitzari dagozkien bektoreak edukita, eredu honek kontsultaren eta dokumentuaren arteko antzekotasuna kalkulatuko du, kontsultaren bektorea eta dokumentu bakoitzarena alderatuz banan-banan. Alderaketa hau bi bektoreen arteko angeluaren kosinua izan daiteke, adibidez. Hortaz, dokumentuak bi multzotan —adierazgarriak eta ezadierazgarriak— banatu beharrean, dokumentuen ranking bat egiten du, eta ranking horretan partzialki bakarrik parekatzen ez diren dokumentuak ere egongo dira.

Termino-haztapena

Esan dugun moduan, bektoreko elementu bakoitzak pisu jakin bat izango du. Pisu hauek estatistiketan oinarrituta kalkulatu ohi dira. Esate baterako, oso ezagunak dira *tf* eta *idf* pisuak. *tf* delakoa ingelesezko *term frequency*tik dator, eta dokumentu batean termino batek duen agerpen kopurua adierazten du. *idf* (*inverse document frequency*) balioak, hau da, dokumentu-maiztasunaren alderantzikatua deituko diogunak, terminoa bildumako zenbat dokumentutan agertzen den kontatu eta balio horren alderantzikatua adierazten du: *idf* altua izango du bilduma osoan agerpen gutxi dituen terminoak, eta, alderantziz, baxua izango du ohikoa den terminoak [8, 9]. Termino bati pisua esleitzeko oso erabilia da *tf-idf* neurketa, *tf* eta *idf* balioen biderkadura dena. Termino-haztapenerako (*term weighting*) erabili ohi diren bi balio hauek ez dira bektore-espazioaren ereduari hertsiki lotutako neurriak. Hauen erabilera oso zabala da jarraian ikusiko ditugun beste IB eredu gehienetan ere.

2.3. **Probabilitate-eredu klasikoak**

Maron eta Kuhnsek [10] ekarri zuten IBaren arlora adierazgarritasunaren kontzeptua, honakoa adieraziz: IB sistema batek dokumentu bakoitzari kontsultarekiko duen adierazgarritasun-probabilitate bat esleitu eta probabilitate horien arabera ordenatu beharko lituzke dokumentuak. Ideia honetan oinarritzen da egungo probabilitate-eredu guztien jatorrian dagoen *probability ranking principle* deiturikoa [11].

Printzipio honetan ez da zehazten adierazgarritasuna nola estimatu behar den. Hori dela eta, printzipio honetan oinarriturik hainbat eta hainbat probabilitate-eredu garatu dira: *binary independence model* izenekoa [12, 13], sare bayestarrak [14], BM25 [15], honen aldaera bat den BM25F [16], *Divergence From Randomness* (DFR) [17], eta hizkuntza-ereduak.

2.4. **Hizkuntza-ereduak**

IB probabilitate-eredu klasikoetatik eratorria da hizkuntza-ereduetan oinarritzen dena. Hizkuntza-eredua (*language model*) hitz-segidei probabilitate-banaketak esleitzen dizkien probabilitate-eredu bat da. Hizkuntzaren prozesamenduan asko erabiltzen da, besteak beste hizketaren ezagutzarako eta itzulpen automatikorako. Ponte eta Croft izan ziren lehenak hizkuntza-ereduak IB sistema batean txertatzeko proposamena egin zutenak 1998an [18]. Honako hau da eredu honen oinarrian dagoen ideia: dokumentu baten hizkuntza-eredutik kontsulta sor-badaiteke, dokumentu hori kontsulta horretarako egokia izango da, kontsultako terminoak dokumentuan agertzen badira gertatuko baita hori. Eredu jakin hori *Query Likelihood* (QL) edo kontsulta-egiantza izenez ezagutzen da.

Hori da IBrako hizkuntza-ereduen hurbilpenik oinarrizkoena, baina badira honako beste aukerak ere: *document likelihood* eredua [19], eta inferentzia-sarea eta hizkuntza-ereduak konbinatzen dituena[3] [20].

Berreskurapen-ereduak ugariak eta askotarikoak dira, eta esperimentu anitz egin badira ere, ez dago garbi zein den zein baino hobea [5]. Hiemstrak dio [2] eredu batzuk egokiagoak direla IBko aplikazio jakin batzuetarako, eta beste eredu batzuk beste aplikazio batzuetarako.

3. **PAREKATZE ARAZOA**

Ereduak eredu, kontsulten eta dokumentuen arteko bat etortzea neurtzean, kontsultan erabilitako hitzen eta dokumentu-indizeetan erabilita-

[3] *Indri* izeneko IB sistema azken eredu berritzaile honetan oinarritzen da: http://www.lemurproject.org/indri/

koen parekatzea prozesatzen da. Kontsulta eta dokumentuetan erabiltzen den hizkuntzak, edozein delarik ere, honako bi ezaugarri hauek izango ditu behintzat:

— Aberatsa: ideia edo gauza bat adierazteko hitz edo esamolde bat baino gehiago erabil ditzakegu.
— Anbiguoa: hitz batek hainbat interpretazio ditu agertzen den testuinguruaren arabera.

Hizkuntzen aberastasun hori nabarmen islatu da hainbat ikerketatan: «bi pertsonak gauza bera deskribatzeko termino bera erabiltzeko probabilitatea % 20 baino txikiagoa da» [21]; «objektu bat emanik bi pertsonak hitz bera esleitzeko probabilitatea % 7-18 bitartekoa da» [22]. Anbiguotasunari dagokionez, esaterako, WordNet ezagutza-base lexikalean dauden 155.287 hitzetatik % 17,3 polisemikoak (adiera bat baino gehiago dituzten hitzak) dira, eta aditz eta izen bakoitzaren batez besteko adiera kopurua 2,17 eta 1,24 da, hurrenez hurren[4].

Hizkuntza baten aberastasun eta anbiguotasun horiekin lotutako bi fenomeno linguistikok, sinonimia eta polisemia, oraindik ere arazo bat dira IB sistemetarako. Bi fenomeno hauek desberdin eragiten diote berreskurapen prozesuari. Sinonimia dela eta, zaila izango da kontsultako ideia bera beste hitz batzuekin adierazia duten dokumentuak berreskuratzea; alegia, nahi baino dokumentu gutxiago berreskuratzea ekar lezake. Polisemiaren eraginez, berriz, zarata sartzen da dokumentu berreskuratuen zerrendan, dokumentu ezadierazgarriak berreskuratzen direlako. Dokumentu ezadierazgarri hauek berreskuratzearen arrazoia honakoa da: dokumentu hauek kontsultako terminoak dituzte, baina kontsultan eta dokumentuan termino hauek duten esanahia ez da berdina. Adibide batzuekin argiago ikusiko dugu oraintxe azaldu berri duguna.

3. irudiko adibide bakoitzean kontsulta (K) eta berari dagokion dokumentu adierazgarria (D) azaltzen da. 3a adibideko kontsultako hitzetatik bakarra (*MEMS*) agertzen da dokumentuan. Alabaina, kontsultako beste hitzarekin zerikusia duten hitz batzuk agertzen dira dokumentuan (*mikroi* eta *diametro*) eta horiek egiten dute dokumentua adierazgarri. Antzeko zerbait ikus dezakegu 3b adibidean ere; kontsultan agertzen diren *koneju*, *hiltze* eta *arrazoi* hitzak ez, baina lehenengo bien sinonimo (*untxi* eta *heriotza*) eta azkenarekin zerikusia duen hitza (*eragile*) azaltzen dira dokumentuan. Gizakiok kontsulta eta dokumentu horiek irakurri orduko konturatzen gara dokumentu horiek badutela kontsulta horietan eskatzen den informazioa. Baina, kontsulta eta dokumentuetan agertzen diren terminoen karaktere-

[4] Datu hauek hemendik hartu dira: http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

K: MEMS tamaina

D: MEMSek (egitura mikro-elektromekanikoek) 60 eta 70 mikroi-ko diametro a (giza ileak baino txikiagoa) dute eta, teorian, oso zeregin zehatzak izan ditzakete esparru desberdinetan (medikuntza-tresnetan, ingurugiro-kontrolean, autoen airbagaren sentsore gisa...).

(a)

K: koneju hiltze arrazoi nagusi

D: Mixomatosiaren kasua bereziki aipagarria da, UGHren ondoren untxi en heriotz eragile **nagusi**a baita ziurrenik Espainian. Beste herrialde batzuetan mixomatosiarekiko jarkikortasun genetiko altuko populazio lokalak topatu dira; alabaina, osasun-kontrolik gabeko lekualdaketa masiboek animalia-talde horien existentzia eta zaintza baldintza dezakete.

(b)

3. irudia. Sinonimia dela eta, kontsulta (K) eta dokumentuaren (D) arteko parekatze arazoaren bi adibide.

segiden konparatze soil bat besterik egiten ez duen IB sistema batek ez lituzke, ziurrenik, dokumentu adierazgarri horiek berreskuratuko.

Aurrera jarraituz, polisemiarekin zer gertatzen den ikusiko dugu. Jo dezagun 4. irudiko kontsulta egiten duela norbaitek, zuhaitzen ekoizpenean zuhaitzen adarrek eta begiek elkarrekin duten eragina jakin nahirik. Kontsulta horretako hainbat termino polisemikoak dira. Adibidez, *zuhaitz* hitzak esanahi bat baino gehiago ditu: (i) landarea eta (ii) egitura edo eskema bat adierazteko erabiltzen den irudia. *adar* hitza ere polisemikoa da eta hauek dira, besteak beste, bere adieretako batzuk: (i) animalia batzuek dutena, (ii) zuhaitzen enborretik ateratzen dena, (iii) musika tresna eta (iv) bidearen, zientzien, eta abarren zatia. Beste horrenbeste gertatzen da *begi* hitzarekin, besteak beste, (i) ikusmenaren organoa, (ii) landareen puja-hasikina edo (iii) zenbait gaztatan aurki dezakegun zuloa adieraz baitezake. Kontsulta hori erabiliz, IB sistema arrunt batek 5. irudian agertzen diren dokumentu horiek berreskuratzen ditu, beste batzuen artean, kontsultako hainbat hitz agertzen direlako (letra lodiz idatzitakoak). Dokumentu horiek irakurtzen baditugu, berehala konturatuko gara dokumentuak ez direla adierazgarriak kontsulta horretarako, kontsultako hitzek beste esanahi bat dutelako. Kontsultako *zuhaitz* hitzaren adiera landarearena da, eta 5a dokumentuan zuhaitz genealogikoarena. *adar* hitzaren kasuan, kontsultan zuhaitz baten adarra adierazi nahi du, baina, dokumentuetan eskema grafiko baten adarra eta gorputz-adarra (5a), eta sailkapen edo zientziaren ada-

rra (5b eta 5c) adierazten da. Antzeko zerbait gertatzen da *begi* hitzarekin ere, kontsultan zuhaitz baten begia adierazi nahi baitu, baina, lehenengo bi dokumentuetan ikusmen-organoari egiten dio erreferentzia, eta azkenekoan begi-bistako espresio moduan erabiltzen da.

zuhaitz ekoizpen **adar** **begi**

4. irudia. Hainbat termino polisemiko dituen kontsulta baten adibidea.

LURREKO UGAZTUNEN ETA BALEEN ARTEKO ZUBIA

Aurkikuntza berriak baleen eta lehorreko ugaztunen arteko zubi dira; garrantzi handikoak beraz. Besteak beste, ugaztunen **zuhaitz** geneologikoan bi **adar**ren arteko adabegi berria finkatzen dutelako. (...) Haragijaleen hortzak dituzte (nahiz eta benetako letaginak ez dituzten) eta txakurrek baino buztan luze eta indartsuagoa, mutur luzeagoa eta **begi** txikiagoak. Biek baleek soilik dituzten hezur bereziak dituzte belarrietan. (...) Pakistanen aurkitutako duela 47 milioi urteko beste fosil batzuek ere ungulatuen orkatilako hezurra dute, baina soin-**adar**rak uretara moldatuagoak dituzte. Baleek duten itxura dela eta (lurreko ugaztunetatik hain desberdina, arraun-moduko hankekin eta muturra buruaren goialdean dutela), oso zaila izan da **zuhaitz** ebolutiboan toki egokian jartzea. (...)

(a)

MARTIALIS: INURRIEN AZPIFAMILIA BERRI BATEKO LEHENA

Inurri arraro bat aurkitu dute Brasilgo basoetan. (...) Inurria espezie berri batekoa izateaz gain, ordea, sailkapen taxonomikoaren **adar** berri bat zabaldu du. Genero ezezaguneko inurria da, eta, are gehiago, azpifamilia baten lehen kidetzat hartu dute adituek. (...) Gainera, **begi** gabekoa da (lurpeko animalia, beraz), baraila luze eta estuak ditu, eta atzeko hanka-parea oso gutxi garatua du, aurrekoekin alderatuta. (...)

(b)

SINETSI BEHARREKO ZERBAIT

(...) Zientziaren **adar** guztiek erabiltzen dute, gainera, berezko axioma-multzo bat. (...)Axioma batzuk **begi**-bistakoak dira; adibidez, 1=1 axioma bat da. Ezin da frogatu 1=1 dela, baina egiatzat hartzen dugu. Oso-oso **begi**-bistakoak iruditu arren, sinplekeria horiek definitzea oso garrantzitsua da, axioma teorien oinarria delako; eta oinarri horietako bat ez bada zuzena, horren gainean eraikitako zientzia guztia zakarretara bota beharrekoa da. **Adar** batzuetan, gainera, eztabaida sutsua da(...)

(c)

5. irudia. Polisemia dela eta, aurreko adibideko kontsultarentzat berreskuratutako dokumentu ezadierazgarrietako batzuk.

Adibide horiek garbi erakutsi digute dokumentu bat adierazgarria den edo ez erabakitzerakoan ezin dela irizpide bakartzat hartu kontsultako hitzak bere horretan agertzea. Horrenbestez, ondoriozta dezakegu hitz horien esanahiak kontuan ez hartzeak arazoak sortzen dituela.

Parekatze arazo horren dimentsioa erakutsi nahi izan dute zenbait egilek. Hala, Swansonek dioenez [23], O'Connor [24] hitzen maiztasun eta indizeen arteko lotura aztertzen ari zela, beste gauza batzuen artean, honetaz konturatu zen: *toxicity* gaitzat zuten 23 dokumentutatik 11 dokumentutan *toxi* errodun hitzik ez zen agertzen. Argi dago, dokumentu hauek berreskuratzeko kontsulta egokiak egitea zaila litzatekeela, gaia *toxicity* izanagatik ere, litekeena baitzen 11 dokumentu horiek ez lortzea, kontsultan hitz hori erabiliz gero.

Lan berriago batean, kontsultako eta dokumentu adierazgarrietako terminoen arteko parekatze eza gertatzen dela ziurtatzeko, TRECeko datu multzo baten gainean, kontsultetako eta dokumentu adierazgarrietako terminoen arteko gainjartzea zenbatekoa zen kalkulatu zuten, eta hauek izan ziren emaitzak [25]: dokumentu adierazgarrien artean % 35,5etan kontsultako termino guztiak agertzen ziren, eta % 13,5 dokumentu adierazgarritan ez zen kontsultako terminorik agertzen. Hortaz, kontsultako terminoen parekatze soil baten bidez ezin ziren berreskuratu azken dokumentu horiek.

4. SEMANTIKA LEXIKALA PAREKATZE ARAZOARI AURRE EGITEKO

Sinonimiak eta polisemiak sortutako arazo horiek izan dira lan hau egitera bultzatu gaituztenak. Hizkuntzaren prozesamenduaren (HP; ingelesez *Natural Language Processing* edo NLP) arloko hitzen adiera-desanbiguazioaz (HAD) [26] eta ahaidetasun semantikoaz [27] baliatu gara parekatze arazoei aurre egiteko.

HPko teknika horiek IB metodoekin konbinatu ahal izateko, hedapena (*expansion*) deritzon teknika erabili dugu. Teknika hau kontsultei (kontsulta-hedapena edo *query expansion*) edo dokumentuei (dokumentu-hedapena edo *document expansion*) aplikatzen zaie, eta, hitz gutxitan esanda, kontsultei edo dokumentuei hitz berriak gehitzean datza. Gure kasuan, gehitutako hitz berri horiek antzekotasun edo ahaidetasun semantikoren bat izango dute kontsultan edo dokumentuan agertzen diren hitzekin.

Adibideetara itzuliz, 3a irudiko kontsulta semantikoki aztertuz, neurriekin zerikusia duten hitzekin hedatuko genuke hasierako kontsulta hori, esaterako, *metro*, *hazbete*, *mikroi*, *handi* edo *diametro*. Hedapeneko hitzak kontuan hartzen dituen IB sistema batek, hedatutako kontsulta eta dokumentua parekatzean, adibideko dokumentu hori amaierako dokumentuen

zerrendan goragoko postuan kokatuko luke, hitz komunen kopuru altuagoa dela eta.

HADaren edo ahaidetasun semantikoaren bidez polisemia ebazteko aukeretako bat da kontsulta eta dokumentuak desanbiguatzea, alegia, HAD sistema baten bidez hitz guztiak dagokien adierekin etiketatzea.

Horrela, IB sistemak hitzak parekatu beharrean, adierak pareka ditzake. 4. eta 5. adibideetako kontsulta eta dokumentuetako hitzak desanbiguatuko bagenitu, eta, ondoren, adierak kontuan hartuta egingo bagenu bilaketa, sistemak ez lituzke dokumentu horiek aukeratuko, adierak ez datozelako bat. Aldiz, 6. irudiko dokumentua, adierazgarria dena 4. kontsultarako, adierazgarritzat hartuko luke, adierak bat datozelako.

PINUEN HOBEKUNTZA GENETIKOA.
Neiker zentro teknologikoa pinus radiata-ren **ekoizpen**a handitzea helburu duen proiektuan ari da lanean 1982. urtetik. Hobekuntza genetikoko programa 1982an jarri zen abian. Lehenik, Euskal Herriko pinu landaketak miatu eta pinurik onenak aukeratu ziren, 80 bat guztira. Pinu horiek **ekoizpen**-ezaugarriak kontuan izanda aukeratu ziren, tamainaz handiagoak zirelako edo **adar** gutxiago zituztelako. Izan ere, **adar** gutxiago izanda, egurrak **begi** edo korapilo gutxiago izaten ditu, eta, ondorioz, kalitate handiagoa. (. . .)

6. irudia. 4. kontsultarentzako dokumentu adierazgarria.

Beste aukera bat da ahaidetasun semantikoren bat duten hitzekin hedapena egitea, HADa esplizituki egin gabe. Hedapenaren ondoren lortuko den kontsultaren edo dokumentuaren errepresentazioak benetako esanahiaren zantzu gehiago edukiko du, semantikoki aberatsagoa izango da, eta, hortaz, parekatzeko hitz gehiago izango ditu. Nabarmentzekoa da ahaide semantikotzat hartzen ditugula, sinonimoez gain, elkarren artean *nolabaiteko* harreman edo zerikusi semantikoa duten hitzak. Ikuspegi zabalago horrek areagotu egiten ditu bilaketarako aukerak.

Laburbilduz, adibideen bidez ohar gaitezke kontsulta eta dokumentuak hitz-segida moduan ikusi beharrean, hitz horien esanahia ere kontuan hartuz gero, handitu egiten dela berreskurapenean arrakasta lortzeko aukera. Garatu dugun metodoan *hitzen adiera-desanbiguazioa eta ahaidetasun semantikoa erabili ditugu kontsulta eta dokumentuak hobeto ulertzeko. Hedapena erabiliz hitz berriak lortu eta informazio hori IB sisteman txertatuz,* saiatu gara berreskurapen prozesutik lortzen diren dokumentu adierazgarri gehiago goragoko posizioetan kokatzen [28].

Kontuan hartu behar da, hala ere, hurbilpen honek badituela mugak eta arriskuak; izan ere hedapen-hitzak gehitzeak parekatze desegokietarako aukera ere areagotu dezake, hitz gehitu horiek ondo aukeratzen ez badira.

Horretan datza metodoaren arrakasta: hedapen-hitzek mesedea egin behar dute, ez kaltea.

5. HEDAPEN-TEKNIKAREN APLIKAZIOA

Deskribatu dugun hedapen-teknika ebaluatu ahal izateko, esperimentazio-ingurunea antolatu, eta bertan hainbat proba egin ditugu.

Oro har, zeregin neketsua eta kostu handikoa da esperimentazio-ingurunea ondo osatzea, besteak beste eskuzko lan handia eskatzen duelako. Horregatik, ingeleserako eta beste zenbait hizkuntza handitarako izan ezik, nekez aurkitzen da ingurune egokirik bestelako hizkuntzetarako.

Guk ingelesera jo dugu gure teknikaren portaera ebaluatu ahal izateko. Hiru datu multzo baliatu ditugu, hirurak IBaren arloko nazioarteko komunitatean erabiliak.

— Robust-WSD: *Cross Language Evaluation Forum*eko (CLEF) Robust-WSD atazan erabilitakoa [29].
— Yahoo!: datu multzo hau *Yahoo! Answers* webgunearen iraulketa baten azpimultzoa.
— ResPubliQA: CLEFeko *Multilingual Question Answering* atazako ResPubliQA ariketarako prestatutakoa.

1. taulan jarri ditugu datu multzo horien hainbat xehetasun (dokumentu eta kontsulta kopuruak eta luzerak hitzetan neurtuak).

1. taula. Datu multzo bakoitzaren xehetasunak: dokumentu kopurua,dokumentuen batezbesteko luzera (hitzak), entrenamendurako eta testerako kontsulten kopurua, eta kontsulten batezbesteko luzera (hitzak).

datu multzoa	dokumentuak		entrenamendu-kontsultak		test-kontsultak	
	#	luzera	#	luzera	#	luzera
Robust	166.754	532	150	8,37	160	8,64
Yahoo!	89.610	104	1.000	11,32	30.000	11,25
ResPubliQA	1.379.011	20	100	10,22	500	10,71

Dokumentu kopuruak edo luzerak begiratzen baditugu, ikusten da izaera desberdineko datu multzoak direla.

Datu multzo guztietan 2 kontsulta-bilduma dauzkagu: entrenamendurako bildumak eta testerako bildumak. Izan ere, guk erabili ditugun algo-

ritmo guztiek dituzte doitu beharreko parametroak, eta hobe da parametro-doikuntza entrenamendu-bilduman egitea, zeregin horretarako test-bilduma erabili gabe [4].

Egindako esperimentuen bidez, hasierako hipotesia egiaztatu nahi izan dugu: hedapen semantikoan oinarritutako gure teknikak laguntzen al du parekatze arazoari irtenbide egokiagoa ematen? Edo, galdera hori birformulatuz, gure teknikak hobetzen al du IB sistemaren errendimendua?

Askotariko esperimentuak egin ditugu, datu multzo, aldagai, aldaera eta parametro asko ukituz, eta ezin da proba guztietarako balio duen erantzun kategoriko eta bakarra eman. Baina, xehetasunak gorabehera, esan dezakegu oro har gure teknikaren aplikazioak IB sistemaren errendimendua hobetzera egiten duela, eta oso portaera sendoa duela hiru datu multzoetan.

Bereziki azpimarratzekoa da Yahoo! datu multzoaren gainean gure hedapen-teknikek izandako portaera: bai dokumentuak hedatuz, baita kontsultak hedatuz ere, estatistikoki esanguratsuak diren hobekuntzak lortu ditugu eraginkortasun-neurri guztietan.

Bestalde, gure hedapen-teknikak kontsulta-hedapenaren arloan ezaguna den *Pseudo-relevance feedback* (PRF) [4] metodoarekin konparatu ditugu. Kontsulta bakoitzaren emaitza banaka aztertuta, ikusi dugu PRFa eta gure hedapen-teknikak osagarriak direla. Izan ere, PRFak emaitza hobeak lortzen ditu kontsulta errazetan, eta, gure hedapen-teknikak aldiz, hobeak dira kontsulta zailenetan. Horren adibide da, esaterako, ResPubliQA datu multzoko ondorengo kontsulta hau: «*What is the lowest speed in miles per hour which can be shown on a speedometer?*». Kontsulta horrentzako hedapenik gabeko metodoaren eta PRFaren MRR[5] emaitza 0,333koa da. Aldiz, kontsultaren hedapena eginez, MRR=1 lortu dugu, alegia, dokumentu adierazgarria lehenengo postuan lortu dugu. Agian hedapenean *vehicle* eta *distance* hitzak lortu ditugulako topatu dugu dokumentu adierazgarri hori. [30]ean ikus daitezke emaitzen gainean egindako gainontzeko analisiak.

Horrenbestez, hasierako hipotesia berresteko moduan gaude: bai, hedapen semantikoan oinarritutako gure teknikak balio du IB sistemen errendimendua hobetzeko.

5.1. **Euskarazko dokumentuen gainean**

Ingelesezko esperimentazio-ingurunea erabili dugu gure sistemaren portaera aztertzeko eta ebaluatzeko. Baina euskarazko testuinguruan errendimendu-probak egin nahi izanez gero, zoritxarrez, ez dugu gaurdaino ho-

[5] MRR (*Mean Reciprocal Rank*) neurriak lehen dokumentu adierazgarria zein posiziotan berreskuratzen den hartzen du kontuan.

rrelako ingurunerik. Dokumentu bildumak, bai, baditugu eta galderak ere eskura daitezke. Ez dugu, ordea, galdera bakoitzeko dokumentu adierazgarrien bilduma zehazturik.

Hala ere, ikusirik ingeleserako portaera ona erakutsi duela, euskarazko testuinguru batera ekarri dugu gure sistema, Berbatek[6] proiektuaren barruan.

On-line erabilpenerako IB sistema bat inplementatu da Zientzia eta Teknologia gaiak biltzen dituen dokumentu bilduma batean[7]. Bilduma hori Elhuyarrek eskainia da, eta bertan biltzen dira aldizkariko testuak eta irudiak, *Teknopolis* telebista-programako bideoak eta *Norteko Ferrokarrilla* irratsaioko audioak. Esan behar da corpus horrek izaera multimedia duen arren, IB sistemak testuen gaineko bilaketa egiten duela, hau da, irudien, bideoen eta audioen idatzizko transkripzioak erabiltzen direla.

MG4J izeneko oinarrizko bilaketa-motor baten gainean eraikia da IB sistema. Sistema horrek artearen egoerako emaitzak eskaintzen ditu, eta, gainera, badu dokumentuen hedapenaren bidez sortutako indize osagarriak integratzeko aukera ere [31].

6. **ONDORIOAK**

Parekatze arazoa IB sistemen gaur egungo erronketako bat da, beren errendimenduan eragin handia baitu kontsultako eta dokumentuetako hitzen bat-etortzea ondo kudeatzeak. Interes eta gaurkotasun handiko gaia izanik, hurbilpen bat proposatu dugu, hizkuntzaren prozesamenduko baliabideak eta teknikak erabiliz arazoari aurre egiteko. Zehatzago esanda, kontsulten eta dokumentuen hedapena ustiatu dugu, horretarako bi teknika hauek baliatuz: hitzen adiera-desanbiguazioa eta ahaidetasun semantikoa.

Hedapen-prozesu bat proposatu dugu, kontsultako eta dokumentuetako hitzen sinonimoak eta bestelako ahaide semantikoak lortze aldera. Hedapenetik lortutako hitz horiek IB sistemaren prozesuan txertatu eta ustiatzeko modu eraginkor bat azaltzen dugu.

Hiru datu multzotan egindako esperimentuek eta analisiek erakusten dute proposatutako metodoak parekatze arazoari aurre egiteko balio duela eta, ondorioz, baita IB sistemaren eraginkortasuna hobetzeko ere. Ikusirik ingeleserako hiru datu multzotan emaitza onak lortu direla, euskarazko dokumentu bilduma baten gaineko bilaketak egiteko ere inplementatu dugu.

---

[6] http://berbatek.com/

[7] http://ixa2.si.ehu.es/BerbatekDemo/bilatu

7. **ESKER ONAK**

Lan honek Eusko Jaurlaritzaren babesa jaso du Ber2tek (IE12-333) izeneko proiektuan.

8. **BIBLIOGRAFIA**

[1] MOOERS, C.N. 1950. «Information retrieval viewed as temporal signaling». *Proceedings of the International Congress of Mathematicians*.

[2] HIEMSTRA, D. 2009. «Information retrieval: searching in the 21st century». *Chapter Information Retrieval Models, 119. John Wiley & Sons, Ltd*. ISBN 9780470033647.

[3] BUSH, V. 1945. «As we may think». *The Atlantic Monthly,* **176(1),** 101-108.

[4] MANNING, C.D., RAGHAVAN, P. eta SCHÜTZE, H. 2009. «An introduction to information retrieval». Cambridge University Press, UK.

[5] BAEZA-YATES, R. eta RIBEIRO-NETO, B. 2011. «Modern information retrieval - the concepts and technology behind search». Second edition. Pearson Education Ltd., Harlow, England. ISBN 978-0-321-41691-9. http://www.mir2ed.org/.

[6] LUHN, H.P. 1957. «A statistical approach to mechanized encoding and searching of literary information». *IBM Journal of Research and Development,* **1(4)**, 309-317. ISSN 0018-8646.

[7] SALTON, G. 1971. «The SMART retrieval system - Experiments in automatic document processing». Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[8] ROBERTSON, S. 2004. «Understanding inverse document frequency: On theoretical arguments for IDF». *Journal of Documentation*, **60(5)**, 503-520.

[9] JONES, K.S. 1972. «A statistical interpretation of term specificity and its application in retrieval». *Journal of Documentation*, **28(1)**, 11-21.

[10] MARON, M.E. eta KUHNS, J.L. 1960. «On relevance, probabilistic indexing and information retrieval». *J. ACM,* **7(3)**, 216-244. ISSN 0004-5411.

[11] ROBERTSON, S.E. 1977. «The probability ranking principle in IR». *Journal of Documentation*, **33(4)**, 294-304.

[12] ROBERTSON, S.E. eta JONES K.S. 1976. «Relevance weighting of search terms». *JASIS*, **27(3)**, 129-146. ISSN 1097-4571.

[13] VAN RIJSBERGEN, C.J. 1979. «Information retrieval». Butterworths, London, 2nd edition.

[14] TURTLE, H. eta CROFT, W.B. 1991. «Evaluation of an inference network-based retrieval model». *ACM Trans. Inf. Syst*, **9**, 187-222. ISSN 1046-8188.

[15] ROBERTSON, S.E. eta WALKER, S. 1994. «Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval». *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '94*, 232-241, New York, NY, USA. Springer-Verlag New York, Inc. ISBN 0-387-19889-X.

[16] ROBERTSON, S., ZARAGOZA, H. eta TAYLOR, M. 2004. «Simple BM25 extension to multiple weighted fields». *Proceedings of the thirteenth ACM international conference on Information and knowledge management, CIKM '04*, 42-49. New York, USA. ACM. ISBN 1-58-113-874-1.

[17] AMATI, G. eta VAN RIJSBERGEN, C.J. 2002. «Probabilistic models of information retrieval based on measuring the divergence from randomness». *ACM Trans. Inf. Syst.*, **20(4)**, 357-389.

[18] PONTE, J.M. eta CROFT, W.B. 1998. «A language modeling approach to information retrieval». *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98*, 275-281, New York, USA. ACM. ISBN 1-58113-015-5.

[19] LAVRENKO, V. eta CROFT, W.B. 2001. «Relevance based language models». *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '01,* 120-127, New York, USA. ACM. ISBN 1-58113-331-6.

[20] METZLER, D. eta CROFT, W.B. 2004. «Combining the language model and inference network approaches to retrieval». *Inf. Process. Manage.*, **40**, 735-750. ISSN 0306-4573.

[21] BATES M.J. 1986. «Subject access in online catalog: a design model». *Journal of The American Society for Information Science*, 357-376.

[22] FURNAS, G.W., LANDAUER, T.K., GOMEZ, L.M. eta DUMAIS, S.T. 1987. «The vocabulary problem in human-system communication». *Communications of the ACM*, **30(11)**, 964-971.

[23] SWANSON, D.R. 1988. «Historical note: information retrieval and the future of an illusion». *JASIS*, **39(2)**, 92-98.

[24] O'CONNOR, J. 1961. «Some remarks on mechanized indexing and some small-scale empirical results». *Machine Indexing: Progress and Problems*, 262-279. The American University.

[25] MULLER, C. eta GUREVYCH, I. 2009. «A study on the semantic relatedness of query and document terms in information retrieval». *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 3 lib. of EMNLP '09*, 1338-1347, Stroudsburg, PA, USA. Association for Computational Linguistics.

[26] AGIRRE E. eta EDMONDS P. (editors). 2006. «Word sense disambiguation: algorithms and applications», *Text, Speech and Language Technology Series, 33 lib*. Springer.

[27] BUDANITSKY, A. eta HIRST, G. 2006. «Evaluating WordNet-based measures of lexical semantic relatedness». *Computational Linguistics*, **32**, 13-47. ISSN 0891-2017.

[28] OTEGI, A. 2012. «Hedapena informazioaren berreskurapenean: hitzen adiera-desanbiguazioaren eta antzekotasun semantikoaren ekarpenak». Doktoretza-tesia.

[29] AGIRRE, E., DI NUNZIO, G.M., MANDL, T. eta OTEGI, A. 2010. «CLEF 2009 Ad Hoc Track Overview: Robust-WSD Task». *Multilingual Informa-*

tion Access Evaluation I. Text Retrieval Experiments, CLEF 2009, *6241 lib. of Lecture Notes in Computer Science*, 36-49. Springer. ISBN 978-3-642-15753-0.

[30] OTEGI, A., ARREGI, X. eta AGIRRE, E. 2014. «Using knowledge-based relatedness for information retrieval». *Knowledge and Information Systems*. Argitaratzeke.

[31] ROBERTSON, S. eta ZARAGOZA H. 2009. «The probabilistic relevance framework: BM25 and beyond». *Foundations and Trends in Information Retrieval*, **3(4)**, 333-389. ISSN 1554-0669.