EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

Doktorego-tesia

Hedapena informazioaren berreskurapenean: hitzen adiera-desanbiguazioaren eta antzekotasun semantikoaren ekarpenak

Arantxa Otegi Usandizaga

2011

EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

Hedapena informazioaren berreskurapenean: hitzen adiera-desanbiguazioaren eta antzekotasun semantikoaren ekarpenak

Arantxa Otegi Usandizagak Eneko Agirre Bengoa eta Xabier Arregi Iparragirreren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua.

Donostia, 2011ko abendua.

Eskerrak

Atzera begira jarri eta errepaso azkar bat eginez, azken urte hauetan modu batera edo bestera lagundu nautenei eskerrak emateko garaia iritsi zait.

Ikerketa-lan honekin hasi baino lehenago IXA taldean pasatako momentuak giltzarri izan ziren lan hau egiten hasteko. Karrera amaitu nuenean *haunditan* zer izan nahi nuen oso garbi ez baneukan ere, taldean jarraitu nahi nuela banekien, gustura sentitzen bainintzen bertan. Taldeko *txikiena* izateagatik jaso nituen mimoengatik izango zen beharbada. Gero ere halaxe sentitu naiz, gustura. Beraz, urte hauetan guztietan nire taldekide izan zaretenoi, eta bereziki nirekin bulegoa eta bazkalordutan mahaia partekatu duzuenoi, eskerrik asko. Berriak zaretenoi ere bai, oraindik elkarrekin asko egon ez bagara ere, aire berriak beti dira ongi etorriak eta.

Nola ez, taldekideen artean batzuei aipamen berezia egin beharrean naiz. Eneko eta Xabier, Xabier eta Eneko, eskerrik asko hain ondo gidatzeagatik eta *zaintzeagatik*. Ez nuen uste momentu honetara hain *sano* eta lasai iritsiko nintzenik! German, a ti también, eskerrik asko por haber estado ahí como director suplente y ser director titular cuando lo necesitaba. Grafo, HAD eta terminologia kontuekin lagundu didazuenoi ere eskerrak eman nahi dizkizuet. Eta *jibotusa*ren *IndriRunQuery*ak jasan behar izan dituzuenoi zer esan... ba izugarrizko pazientzia izan zenutela!

Bartzelonan eta Amsterdamen egon nintzen bitartean gidatu eta lagundu zenidatenoi (Hugo, Maarten eta Edgar) gracias eta thank you! Bartzelonako EMPEs kuadrilakoak ere ezin ahaztu, moltes gracies!

Unibertsitatetik pixka bat aldenduz, ezin ahaztu etxekoak. Badakit, ez duzue oso ondo ulertzen zertan ari naizen, baina, berdin dio. Behar izan dudanean sumatu dut zuen babesa eta horrekin nahikoa izan dut, eskerrik asko. Ikustek? Abenduan bukatu diat! Baina ze urtetako abenduan? Hori beste kontu bat duk... Etxeko txikitxoak: agobioei aurre egiteko zuekin egotea bezalakorik ez dago!

Koadrilakoei beste hainbeste. Aizue, *liburu potolua* amaitu dut (iufi!) eta hasia naiz bazkari edo afari horretan pentsatzen, eh! Ya^{juuu}uuu!! Beste lagunak ere aipatu nahi ditut. Zuekin honetaz apenas hitz egin dudan, askok ez dakizue zertan aritu naizen ere. Baina berdin dio, asteburuetan zuekin egoteak astelehenetan honetan gogotsuago hasteko balio izan dit. Hortaz, eskerrik asko.

Zuei denei, berriz ere, eskerrik asko!

Esker instituzionalak

Eusko Jaurlaritzako Hezkuntza, Unibertsitate eta Ikerketa Sailari, ikerketalan hau egiteko emandako ikertzaileak prestatzeko bekarengatik (BFI06.281).

Laburpena

Informazioaren berreskurapena (IB) erabiltzaile baten informazio-beharra asetuko duten dokumentuak bilatzean datza. Honela bada, IB sistemak erabiltzaileari dokumentu adierazgarriak, alegia, erabiltzaileak behar duen informazioa eduki dezaketen dokumentuak, topatzen lagunduko dio, beti ere erabiltzaileak egindako kontsultan oinarrituz. Hain ezagunak eta erabiliak diren Google eta Yahoo! bezalako web-bilatzaileak IB sistemen adibide garbiak dira.

IB sistema perfektu batek dokumentu adierazgarriak bakarrik berreskuratu beharko lituzke, eta ez-adierazgarriak baztertu. Alabaina, sistema perfektuak ez dira existitzen. IB sistemek aurre egin behar dien arazo nagusienetako bat kontsulta eta dokumentuen arteko parekatze-arazoa deiturikoa da: dokumentu bat kontsulta batentzako adierazgarria izan daiteke nahiz eta bietan erabilitako hitzak guztiz berdinak ez izan, eta, alderantziz, dokumentu bat ez-adierazgarria izan daiteke kontsulta batentzat nahiz eta termino batzuk komunean eduki. Lehena ideia edo gauza bera adierazteko hitz edo esamolde bat baino gehiago erabili ditzakegulako (sinonimia) gerta daiteke. Bigarrena, berriz, testuinguruaren arabera hainbat interpretazio izan ditzaketen hitzek (anbiguotasuna) eragiten dezakete. Hau kontuan izanik, IB sistema batek dokumentu bat adierazgarri edo ez-adierazgarri bezala sailkatzerakoan kontuan hartzen duen irizpide bakarra kontsultako hitzak egotea (edo ez egotea) denean zaila suerta daiteke dokumentu egokiak topatzea, eta baita adierazgarriak ez direnak baztertzea. Honen aurrean, hitz horien

esanahiak kontuan hartuz gero berreskurapen arrakastatsuago bat egiteko aukera gehiago egongo direla pentsatzea bidezkoa dirudi.

IBaren hastapenetatik gaur arte parekatze-arazoaren inguruan ikerketa-lan dezente egin badira ere, oraindik guztiz ebatzi gabe jarraitzen du, eta bilatzaile askok ez dute aintzat hartzen. Tesi-lan honetan hizkuntzaren prozesamenduaren (HP) bidez arazo hau arintzerik ba ote den aztertu da.

Hitz gutxitan esanda, kontsulten eta dokumentuen hedapena egiten dugu HPko bi teknikaz baliatuz: hitzen adiera-desanbiguazioa eta ahaidetasun semantikoa. Alde batetik, teknika hauetako bakoitzerako hedapen-prozesu bat proposatzen dugu, non kontsulta eta dokumentuetako hitzen sinonimo eta bestelako ahaidetasuna duten hitzak lortuko ditugun. Bestetik, hedapenetik lortutako hitz horiek, kontsulta eta dokumentuetako jatorrizko hitzekin batera, IB sistemaren prozesuan txertatu eta ustiatzeko modu eraginkor bat azaltzen dugu kasu bakoitzerako. Are gehiago, erabiliko dugun hedapen-teknikak kontsulta eta dokumentuak itzultzeko balio duenez, hedapen-teknika hori erabiliz hizkuntza arteko berreskurapenean hobekuntzak lortzen direla erakutsiko dugu.

Hiru datu-multzotan egindako esperimentu eta analisiek erakusten dute tesi-lan honetan proposatutako hedapen-metodoek parekatze-arazoari aurre egiteko balio dutela eta, ondorioz, baita IB sistemaren eraginkortasuna hobetzeko ere.

"Europako Doktoregoa" aipamena lortzeko Euskal Herriko Unibertsitatearen eskakizunei jarraituz, tesi-txosten honen ingelesezko bertsio laburtua ondorengo helbide honetan aurki daiteke:

http://ixa2.si.ehu.es/~jibotusa/tesia/thesis_summary.pdf

Gaien aurkibidea

1 Tesi-lanaren aurkezpen orokorra

1.1 Informazioaren berreskurapena

Lehen aldiz *information retrieval* terminoa erabili zuena Mooers (1950) izan zen, eta honela definitu zuen:

> "Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. It is the finding or discovery process with respect to stored information."

Erabat errotuta gelditu zen termino hori eta geroztik horrelaxe deitu izan zaio, labur esanda, erabiltzaile baten informazio-beharra asetuko duen dokumentu-bilatzeari. Euskaraz informazioaren berreskurapena (aurrerantzean IB) deritzogu arlo honi.

Honela ba, informazioaren berreskurapenerako sistema bat edo IB sistema bat dokumentuetako informazioa biltegiratu eta kudeatzen duen softwareprograma bat da (Hiemstra, 2009). Sistemak erabiltzaileak behar duen informazioa topatzen lagunduko dio, informazio hori eduki dezaketen dokumentuen berri emanaz. Kontuan izan, horrelako sistemek ez dutela informazioa esplizituki itzultzen edo galdera erantzuten, dokumentuak berreskuratu edo iradoki besterik ez dute egiten.

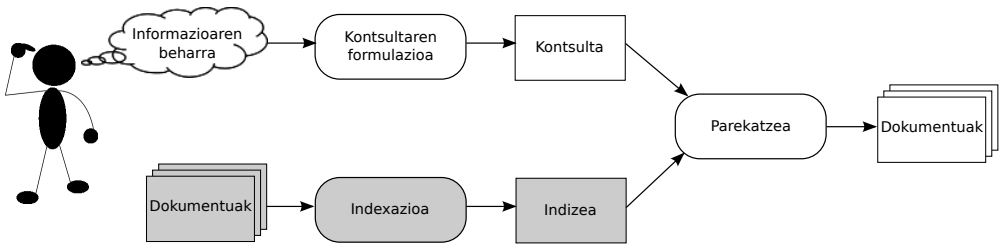
IB sistemen funtzionamendua

IB sistema batek hiru prozesu nagusi gauzatzen ditu (ikus 1.1 irudia):

(i) Indexazioa: dokumentuen errepresentazioa gauzatzen da, indizea(k) sortuz. Indizea bilaketa azkarrak egitea ahalbideratuko duen datu-egitura bat da. Lineaz kanpo (*offline*) egikaritzen da, eta dokumentu-bilduma aldatzen ez bada behintzat, behin egitea nahikoa da.

(ii) Kontsultaren formulazioa: erabiltzailearen informazio-beharra kontsulta batean adierazita jartzen da.

(iii) Parekatzea: kontsulta dokumentuen errepresentazioarekin, indizearekin, parekatzen da. Parekatze honetan dokumentuen azpimultzo bat aukeratzen da.



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

Irteerako azpimultzo horretako dokumentu batzuek, ziur aski, erabiltzailearen informazio-behar hori asetuko dute; dokumentu horiei dokumentu *adierazgarri* deitzen zaie. IB sistema perfektu batek dokumentu adierazgarriak bakarrik berreskuratu beharko lituzke, eta ez-adierazgarriak baztertu. Alabaina, sistema perfektuak ez dira existitzen eta geroago ikusiko ditugu zein diren sistema hauen gabezietako batzuk. Gaur egungo sistemetan ohikoena dokumentu-zerrenda ordenatu bat itzultzea da, zerrendaren hasieran jarriz ustez erabiltzaileari gehien interesatuko zaizkion dokumentuak, alegia, sistemaren ustez adierazgarrienak direnak.

Prozesu horiek konputagailuen bidez guztiz automatikoki egikaritzearen ideia Bush-ek (1945) proposatu zuen lehen aldiz:

> "Consider a future device for individual use, which is a sort of mechanized private file and library. It needs a name, and, to coin one at random, 'memex' will do. A memex is a device in which an

> individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."

Erabilera-esparrua

Bush-en ideia jarraituz, lehenengo IB sistema automatizatuak 50 eta 60ko hamarkadetan sortu ziren. Hasierako urte haietan sistema hauek argitalpen zientifiko eta liburutegietako dokumentuak bilatzeko erabiltzen ziren, batez ere. Bilaketak ez ziren dokumentuen eduki osoaren gainean egiten, baizik eta dokumentuei eskuz esleitutako gako-hitzetan oinarritzen ziren.

Joerak aldatzen joan dira eta gaur egungo egoera guztiz bestelakoa da. Alde batetik, sistema hauen erabilera guztiz zabaldua dago gaur egungo gizartean, azken urteotan konputagailu pertsonalen kopurua handituz eta Internet zabalduz doan heinean, webeko bilatzaileen beharra ere gorantz doalako. Hain ezagunak eta erabiliak diren Google[1] eta Yahoo![2] bezalako webbilatzaileak IB sistemen adibide garbiak dira.

Bilaketak egiteko metodoak

Bilaketak egiteko modua edo, beste modu batera esanda, IB gauzatzeko metodoak ere aldatzen joan dira konputagailuen ahalmena eta biltegiratze-lekua handituz doan heinean. Gaur egungo sistema gehienek dokumentuetan agertzen diren termino guztiak (edo ia guztiak) erabiltzen dituzte bilaketak egiteko, alegia, dokumentuen eduki osoak hartzen dituzte kontuan. Honi ingelesez *full text retrieval* esan ohi zaio. Hala ere, gaur egun ere badira dokumentuen zati jakin batzuetan eta dokumentuei eskuz esleitutako gako-hitzetan oinarritzen diren sistemak. Horren adibide da 70eko hamarkadatik martxan den PubMed bilatzailea[3]. IB sistema honek biomedikuntzako eta osasun-arloko argitalpenak gordetzen dituen MEDLINE datu-basean bilaketak egitea ahalbideratzen du. Datu-base horretan argitalpen bakoitzaren titulu, abstract eta eskuz esleitutako gako-hitzak daude —gako-hitz hauek medikuntzako thesaurus batetik hartutakoak dira.

[1] http://www.google.es/
[2] http://es.yahoo.com/
[3] http://www.ncbi.nlm.nih.gov/pubmed/

Aplikazioak

IB tekniken aplikazio arruntenetakoa bilatzailea da. Bilatzailerik erabilienak web-bilatzaileak badira ere, badira beste batzuk; beste batzuen artean, hauexek:

- Bilatzaile bertikalak: bilaketak domeinu edo gai konkretu batera mugatzen dituzten web-bilatzaile espezializatuak.
- Enpresa-bilatzaileak (*enterprise search*): bilaketak enpresa baten intranetean aurkitzen diren mota desberdinetako dokumentuetan (web-orrialdeak, posta elektronikoa, txostenak, aurkezpenak, kalkulu-orriak, datu-baseak...) egiten dituzten bilatzaileak.
- Mahaigaineko bilatzaileak (*desktop search*): bilaketak konputagailu pertsonalean aurkitzen diren dokumentuetan egiten dituzten bilatzaileak. Kasu honetan, aurrekoan bezala, dokumentu horiek mota askotakoak izango dira.

Horrelako bilatzaileez gain, ordea, badira IBaren aplikazio orokor gehiago ere. Izan ere, testu-bilduma edo bestelako informazio ez-egituratua darabilen edozein aplikaziok, informazio hori antolatu eta bilatu beharko du momenturen batean. Horren adibide dira, esaterako, honako hauek:

- Liburutegi digitalak: edukiak formatu digitalean gordetzen dituzten liburutegiak, ordenagailu bidez atzitzen direnak.
- Informazioaren iragazketarako tresnak (*information filtering systems*) eta hauen artean aurkitzen diren gomendio-sistemak (*recommender systems*). Azken hauek, izenak dioen bezala, erabiltzailearen intereseko izango diren informazio-elementuak (pelikulak, liburuak, musika, ikuskizunak...) gomendatzen dituzten sistemak dira.

Atazak

Hainbat atazatan erabiltzen dira IB teknikak; besteak beste, ondoren zerrendatuko ditugun hauetan, azkenekoa delarik aipagarriena:

- Dokumentu-sailkapena: dokumentu bakoitzaren edukian oinarrituz, dokumentuei etiketa bat esleitzea edo klase konkretu batekoak direla adieraztea.
- Galderak erantzutea (*question answering*): erabiltzaileak lengoaia naturalean egindako galderari erantzun zehatza bilatzea.
- Dokumentu askotan oinarritutako testu-laburtze automatikoa (*multi-document automatic summarization*): gai jakin bati buruzko hainbat

dokumenturen laburpena izango den dokumentu bakar bat lortzea.
- ***Ad hoc***: **erabiltzaileak mahaigaineratutako kontsulta batean adierazitako informazio-beharra asetuko duten dokumentuak bilatzea dokumentu-bilduma batean**. Atazarik arruntena dela esan daiteke eta tesi-lan honetan aztergai izango duguna da.

Bilaketa-objektuak

Gaur egun gero eta ohikoagoa da testua ez diren dokumentuekin (irudiak, bideoak, audio-fitxategiak edo eskaneatutako dokumentuak) lan egitea arlo honetan. Hala ere, hasiera hasieratik gehien erabiltzen diren dokumentuak **testu-dokumentuak** dira. Horregatik, nahiz eta testu-dokumentuekin bakarrik aritu, askotan *informazioaren berreskurapena* termino orokorra erabili ohi da, dokumentuaren berreskurapena (*document retrieval*) edo testuaren berreskurapena (*text retrieval*) esan beharrean. Tesi-txosten honetan ere halaxe egingo dugu aurrerantzean; gure esperimentu guztiak testu-dokumentuekin egin arren, aipatutako hiru terminoak erabiliko ditugu bereizketarik egin gabe.

Ikerketa-gaiak

Ikusi dugun moduan, IB sistema batek erabilera askotarikoak izan ditzake eta urteak joan ahala esparru berrietara zabalduz doanez, ikerketa-gaiak ez dira falta arlo honen inguruan: 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 edo parekatze-arazoa.

Azken gai hori izango da tesi-lan honetan jorratuko duguna. Gehiago zehaztuz, gure ikerketa-gaia hauxe izango da: **ad hoc ataza batean testu-dokumentuak bilatzerakoan izaten den parekatze-arazoa**. Jarraian, gai hau luze eta zabal jorratuko dugu.

1.2 Parekatze-arazoa

IBaren hastapenetatik gaur arte parekatze-arazoaren inguruan ikerketa-lan dezente egin badira ere, oraindik guztiz ebatzi gabe jarraitzen du. Goazen bada ikustera zein den arazoaren iturria.

Eguneroko komunikaziorako darabilgun hizkuntza edozein delarik ere, honako bi ezaugarri hauek behintzat izango ditu:

- aberatsa: ideia edo gauza bat adierazteko hitz edo esamolde bat baino gehiago erabil ditzakegu;
- anbiguoa: hitz batek hainbat interpretazio ditu agertzen den testuinguruaren arabera.

Egindako hainbat ikerketak hizkuntzaren aberastasun hori islatzen dute. Adibidez, Bates-ek (1986) egindako esperimentuetan honakoa ikusi zuen: "bi pertsonak gauza bera deskribatzeko termino bera erabiltzeko probabilitatea % 20 baino txikiagoa da". Furnas *et al.*-ek (1987) ere, horixe berretsi zuten: "objektu bat emanik bi pertsonak hitz bera esleitzeko probabilitatea % 7-18 bitartekoa da". Anbiguotasunari dagokionez, esaterako, WordNet ezagutza-base lexikalean dauden 155.287 hitzetatik 26.896 hitz (% 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 (adiera bakarrekoak kontuan hartu gabe kopuru hauek 3,57 eta 2,79koak dira)[4].

Gauzak horrela izanik, eta IB sistemek kontsulta eta dokumentuak parekatzerakoan hauetan agertzen diren karaktere-kateen arteko parekatze soil bat besterik ez badute egiten behintzat —eta horixe da egiten dutena oinarrizko berreskurapen-ereduek, baita oraintsu arte sistema komertzial gehienek ere—, ez da harritzekoa izango dokumentu egokiak bilatzeko zailtasunak izatea. Besteak beste, hizkuntza batek berezko dituen ezaugarri horiek dira parekatzean zailtasunak eragiten dituztenak. IB sistema batek honako fenomeno linguistiko hauei egin behar die aurre:

- aldaera sintaktikoak
  euria ari duelako etxean gelditu naiz / etxean gelditu naiz euria ari duelako
- aldaera morfologikoak
  abesti / abestiak / abeslari / abestu / abestuko / abesten
- aldaera morfosintaktikoak
  industria-jarduera / jarduera industrial
- aldaera lexikalak: **sinonimia**
  auto / automobil / beribil
- aldaera semantikoak: **polisemia**
  kaiku: euskal jaka / ontzia / inozo, buruarin, ergel
- hizkuntza batetik besterako aldaerak: hizkuntza batean dagoen kontsultaren bidez beste hizkuntza batean dauden dokumentuak berresku-

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

ratzeko aukera dagoenean sortzen direnak.

Esan daiteke aldaera sintaktikoek ez dutela gaur egun gehien erabiltzen diren sistemetan arazo handirik sortzen. Izan ere, *bag of words* deitzen zaion eredua jarraitu ohi da gehienetan, eta eredu honetan, izenak adierazten duen bezala, dokumentuak hitzez betetako zaku bezala ikus daitezke. Hau da, dokumentuan hitzek duten segida erabat galtzen da, eta, hortaz, ordena ez da kontuan hartzen bilaketak egiterakoan. Gaur egungo sistema batzuek aukera ematen dute hitzak ordena jakin batean bilatzeko, baina orduan espresuki adierazi behar da hori.

Aldaera morfologiko eta morfosintaktikoak direla eta, hitz (edo esamolde) baten aldaera asko lor ditzakegu. Bilaketa bat egin nahi dugunean, kontsultan hitz bat jarri arren, bere aldaera guztiak dituzten dokumentuak bilatzea interesgarria da gehienetan; horra aldaera hauek eragiten duten arazoa, parekatze soil batekin kontsultako hitz berberak bakarrik topatuko baitira, eta ez hitz horren antzekoak edo familiakoak. Dena den, arazo hauek nahiko ondo ebazten dira erro-bilatzaile (*stemmer*) edo lematizatzaileekin. Tresna hauek hitz bakoitzaren erroa edo lema zein den esaten dute, eta erro edo lema horiek erabiltzen dira indizeak sortzeko eta bilaketetarako. Horrela, adibidez, euskararako lematizatzailea erabiltzen badugu, abesti eta abestiak hitzak abesti hitzarekin ordezkatuko dira kontsulta eta dokumentuetan; berdina gertatuko da abestu, abestuko eta abesten hitzekin, horien ordez abestu hitza erabiliko da eta.

Sinonimiak eta polisemiak oraindik ere IB sistementzat arazo izaten jarraitzen dute. Bi fenomeno hauek modu desberdinean eragiten diote berreskurapen-prozesuari. Sinonimia dela eta, kontsultako ideia bera beste hitz batzuekin adierazia duten dokumentuak berreskuratzea zaila izango da; alegia, nahi baina dokumentu gutxiago berreskuratzea ekar lezake. Polisemiak, berriz, zarata sartzen du berreskuratutako dokumentu-zerrendan, kontsultako terminoak, baina beste esanahi batekin, dituzten dokumentu ez-adierazgarriak berreskuratzen direlako. Adibide batzuekin argiago ikusiko dugu oraintxe azaldu berri duguna[5].

1.2 irudiko adibide bakoitzean kontsulta (Q) eta honi dagokion dokumentu adierazgarria (D) azaltzen da. 1.2a adibideko kontsultako gako-hitzak —alegia, bilaketarako erabiliko diren terminoak— *fast* (azkar), *tractor* (trak-

[5]Adibide hauek gure esperimentuetan erabili ditugun datu-multzoetatik hartutakoak dira, eta, beraz, ingelesez daude. Honen inguruan xehetasun gehiago 1.4 atalean emango dugu.

Q: How **fast** does a **tractor go**?

D: This Directive shall apply only to **tractors** defined in paragraph 1 which are fitted with pneumatic tyres and which have two axles and a maximum design speed between 6 and 25 kilometres per hour.

(a) ResPubliQA datu-multzoko 96. galdera eta jrc31977L0537/14 dokumentua.

Q: How do you **cook** an **apple pie**?

D: There are many good recipes for **apple pies** but there are also some important things to remember that are usually not in the recipe. That is you should make sure the bottom of the crust will bake as well and not remain soggy. To do this, coat the inside of the crust with butter before adding the filling and place the baking dish on a dark metal pan so the bottom will get more heat.

(b) Yahoo! datu-multzoko 1005121203620 galdera eta erantzuna.

1.2 irudia – Sinonimia dela eta, kontsulta (Q) eta dokumentuaren (D) arteko parekatze-arazoaren bi adibide.

tore) eta *go* (joan, ibili) dira; baina, hauetatik bakarra (*tractor*) agertzen da dokumentuan. Alabaina, kontsultan agertzen diren hitz horiekin erlazionatutako beste hitz batzuk agertzen dira dokumentuan —*speed* (abiadura) eta *kilometres per hour* (kilometro orduko)— eta horiek egiten dute dokumentua adierazgarri. Antzeko zerbait ikus dezakegu 1.2b adibidean ere; kontsultan agertzen den *cook* (janaria prestatu) gako-hitza ez, baina honekin erlazionatutako hainbat hitz —*recipes* (errezeta) eta *bake* (erre, labean egin)— erabiltzen 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-segiden konparatze soil bat besterik egiten ez duen IB sistema batek ez lituzke, ziurrenik, dokumentu adierazgarri horiek berreskuratuko.

Adibide hauek argi uzten dute sinonimia hutsarekin ez dela nahikoa kontsulta eta dokumentu adierazgarrietako terminoen artean egon daitezkeen desberdintasunak gailentzeko, dokumentuetan dauden hitzak kontsultako hitzekin hertsiki erlazionatuak badaude ere, ez baitira sinonimoak (*fast* hitzarekin *speed* eta *kilometres per hour*, eta *cook* hitzarekin *recipes* eta *bake*). Horrela bada, tesi-lan honetan sinonimiaz haratagoko erlazio lexikoak aztertuko dira. Txosten honetan hemendik aurrera, sinonimia hitzaren zentzua

Title: Computer Mouse RSI
Desc: Find documents that report on computer mouse repetitive strain injuries (RSI).
Narr: Relevant documents report injuries that are caused by the continuous use of a computer mouse. Documents proposing ways to avoid repetitive strain injuries (RSI) when using the computer are also relevant.

(a) Robust-WSD datu-multzoko gaia (10.2452/064-AH).

computer **mouse** **rsi** **repetitive** **strain** **injuries**

(b) Sistemari egingo zaion kontsulta (title eta desc erabiliz).

1.3 irudia – Robust-WSD datu-multzoko kontsulta baten adibidea.

zabalduko dugu, eta *antzekotasun edo ahaidetasunen bat duten hitzak* adierazteko erabiliko dugu. Horrela egingo dugu irakurterrazagoa izan dadin.

Aurrera jarraituz, polisemiarekin zer gertatzen den ikusiko dugu. Eman dezagun 1.3b irudiko kontsulta dugula (1.3a irudian ikusten den informazio-beharretik ateratakoa). Kontsulta horretako hainbat termino polisemikoak dira; adibidez, *mouse* eta *strain* hitzek esanahi edo adiera bat baino gehiago dituzte 1.4 irudian ikus daitekeen moduan. Kontsulta hori erabiliz, gure esperimentuetan erabiliko dugun oinarri-lerroko sistema (*baseline system*) batek 1.5 irudian agertzen diren dokumentu horiek berreskuratzen ditu, beste batzuen artean. Dokumentu horiek irakurtzen baditugu, berehala konturatuko gara kontsultako hitzak agertu arren, dokumentu horiek ez direla adierazgarriak kontsulta horrentzako. Sistemak dokumentu hauek adierazgarritzat hartu ditu, esan bezala, dokumentu horietan kontsultako hainbat hitz agertzen direlako (**kolore honetan** idatzitakoak); baina, horietako gako-hitz batzuk beste esanahi bat dute dokumentuetan. Kontsultako *mouse* hitzaren adiera ordenagailuaren saguarena da, eta dokumentu horietan agertzen denean sagu animaliari buruz ari da. *strain* hitzaren kasuan, kontsultan bihurdura edo zaintiratua adierazi nahi du, eta dokumentuetan hitz horrek beste adiera batzuk hartzen ditu: 1.5a dokumentuan animalia mota eta 1.5b dokumentuan musikaren doinu edo melodia.

Dokumentu bat adierazgarri edo ez-adierazgarri bezala sailkatzerakoan kontuan hartzen den irizpide bakarra kontsultako hitzak egotea (edo ez egotea) denean arazoak sor daitezkeela erakusten digute adibide hauek. Alegia, hitz horien esanahiak kontuan ez hartzeak arazoak sor ditzakeela ikusi dugu.

mouse-1: any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails.
mouse-2, shiner, black eye: a swollen bruise caused by a blow to the eye.
mouse-3: person who is quiet or timid.
mouse-4, computer mouse: a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad.

(a)

strain-1: (physics) deformation of a physical body under the action of applied forces.
strain-2, stress: difficulty that causes worry or emotional tension; *she endured the stresses and strains of life*.
strain-3, tune, melody, air, melodic line, line, melodic phrase: a succession of notes forming a distinctive sequence; *she was humming an air from Beethoven*.
strain-4, mental strain, nervous strain: (psychology) nervousness resulting from mental stress; *his responsibilities were a constant strain*.
strain-5, breed, stock: a special variety of domesticated animals within a species; *he experimented on a particular breed of white rats*.
strain-6, form, variant, strain, var.: (biology) a group of organisms within a species that differ in trivial ways from similar groups; *a new strain of microorganisms*.
strain-7: injury to a muscle (often caused by overuse), results in swelling and pain.
strain-8, tenor: the general meaning or substance of an utterance; *although I disagreed with him I could follow the tenor of his argument*.
strain-9, striving, nisus, pains: an effortful attempt to attain a goal.
strain-10, straining: an intense or violent exertion.
strain-11, song: the act of singing; *with a shout and a song they marched up to the gates*.

(b)

1.4 irudia – *mouse* eta *strain* hitzen WordNet 3.0 bertsioko adierak, *izen* kategoriakoak bakarrik.

RESEARCHER ACCUSED OF FAKING DATA;HER STUDY PURPORTED TO USE GENES TO TRANSFER DISEASE RESISTANCE.
(...) Her results were published in the April 25, 1986, issue of the journal Cell in an article co-authored by Nobel laureate David Baltimore. The article "purposed to show that a gene from one **strain** of **mouse** had been transferred to another **strain** of **mouse**, resulting in the latter's production of high levels of antibody molecules it would not normally produce – antibody molecules mimicking the antibody molecules produced by the original **strain**,"investigators said in a written statement. (...) after reviewing scientific evidence and performing a **computerized** statistical analysis that showed the false data was not made up of chance errors (...)

(a) Robust-WSD datu-multzoko LA112694-0025 dokumentua.

SOUNDS: LATEST WORK IS BOWEN'S MOST HIGH-PROFILE; COMPOSER AND PERFORMER OF NEW MUSIC SPENT YEARS WORKING ON THE FRINGES.
Listening to the lilting **strains** of Gene Bowen's new album "The Vermilion Sea"(...) the Nordic-looking Bowen has a few guitars, a synthesizer and the all-important **computer** – his main composing tool – and piles of records and CDs. (...) Three years ago, Bowen began his work-in-progress, creating the raw material on synthesizers and **computers**. (...) "My interests came through guitar music and songwriting coupled with interest in folk and ethnic music, where **repetition** is always so important. **Repetition** and texture are almost more important than (...)

(b) Robust-WSD datu-multzoko LA063094-0099 dokumentua.

1.5 irudia – Polisemia dela eta, aurreko adibideko kontsultarentzat berreskuratutako dokumentu ez-adierazgarrietako batzuk.

1.3 Semantika lexikala parekatze-arazoari aurre egiteko

Sinonimia eta polisemiak sortutako arazo horiek izan dira tesi-lan hau egitera bultzatu gaituztenak. Hizkuntzaren prozesamenduaren (HP; ingelesez *natural language processing* edo NLP) ikuspuntutik helduko diegu arazoei. HParen baitan hainbat arlo daude, eta guk lexiko-semantikan jarriko dugu arreta. Zehatzago esateko, arlo horretan kokatzen diren hitzen adieradesanbiguazioaz (HAD) (Agirre eta Edmonds, 2006) eta ahaidetasun semantikoaz (Budanitsky eta Hirst, 2006) baliatuko gara (3.4 atalean aurkeztuko ditugu prozesu eta teknika hauek).

Aipatutako lexiko-semantikako teknika hauek IB metodoekin konbinatu ahal izateko, *hedapena* (*expansion*) deritzon teknikaz baliatuko gara. Teknika hau kontsultei aplikatzea ohikoagoa bada ere (kontsulta-hedapena edo

query expansion), dokumentuekin ere egin daiteke (dokumentu-hedapena edo *document expansion*), eta, hitz gutxitan esanda, kontsulta edo dokumentuei hitz berriak gehitzean datza. Gure kasuan, gehitutako hitz berri horiek antzekotasun edo ahaidetasun semantikoren bat izango dute kontsulta edo dokumentuan agertzen diren hitzekin.

Lehengo adibideetara itzuliz, 1.2b irudiko kontsulta semantikoki aztertuz, besteak beste, janaria prestatzearekin (*cook*) erlazionatutako hitzekin hedatuko genuke hasierako kontsulta hori; adibidez, janaria prestatzeko modu desberdinak (*bake*, *boil*, *grill*...), edota agian *cooker* edo *recipe* hitzak ere bai. Hedapeneko hitzak modu egokian kontuan hartzen dituen IB sistema batek hedatutako kontsulta eta dokumentua parekatzean komunean dituzten hitz kopurua altua dela ikusiko luke, eta, ondorioz, adibideko dokumentu hori amaierako dokumentuen zerrendan goragoko postuan jarriko luke.

HADaren edo ahaidetasun semantikoaren bidez polisemia ebazteko hainbat modu daude. Aukeretako bat kontsulta eta dokumentuak desanbiguatzea da, alegia, HAD sistema baten bidez hitz guztiak dagokien adierekin etiketatzea. Horrela, IB sistemak hitzak parekatu beharrean, adierak pareka ditzake. Berriz ere adibideekin jarraituz, 1.3 eta 1.5 irudietan ikusi ditugun kontsulta eta dokumentuak jarri ditugu 1.6 irudian, baina oraingo honetan *mouse* eta *strain* hitzak desanbiguatuta daude (hitzaren ondoren dagokion adiera zehaztu dugu). Horrela, argi eta garbi ikusten da 1.6b eta 1.6c dokumentuetako adierak ez datozela bat 1.6a kontsultako adierekin. 1.6d dokumentua, aldiz, adierazgarritzat hartuko da adierak bat datozelako. Adibide hauetan bi hitz bakarrik desanbiguatu ditugu irakurterrazagoa izan dadin. Alabaina, sistemak hitz guztiak desanbiguatu beharko ditu parekaketa ahalik eta zehatzena izan dadin.

Beste aukera bat ahaidetasun semantikoren bat duten hitzekin hedapena egitea da, HADa esplizituki egin gabe. Konputagailua kontsulta edo dokumentu horren atzean dagoen semantika ulertzeko gai bada, hedapenaren ondoren lortuko den kontsulta edo dokumentuaren errepresentazioak benetako esanahiaren zantzu gehiago edukiko du, semantikoki aberatsagoa izango da, eta, hortaz, parekatzeko hitz gehiago izango ditu. Noski, hedapena (oker) egiteak zarata sartzearen arriskua berekin dakar, eta galera/onura arteko oreka bilatu beharko da. Azken adibide horrekin jarraituz, eman dezagun 1.3b kontsulta hedatzen dugula *electronic device, lesion, wellness* eta kontzeptu horien inguruko hitzekin. Modu berean, pentsa dezagun 1.5b dokumentua hedatzen dugula *instrument, singer, vocalist* eta orokorrean musikarekin zerikusia duten hitzekin. Hedatutako kontsulta eta dokumentuaren arteko pa-

computer **mouse**[mouse-4] **rsi repetitive** **strain**[strain-7] **injuries**

(a) Robust-WSD datu-multzoko 10.2452/064-AH kontsulta desanbiguatuta.

RESEARCHER ACCUSED OF FAKING DATA;HER STUDY PURPORTED TO USE GENES TO TRANSFER DISEASE RESISTANCE.
(...) Her results were published in the April 25, 1986, issue of the journal Cell in an article co-authored by Nobel laureate David Baltimore. The article "purposed to show that a gene from one **strain**[strain-5] of **mouse**[mouse-1] had been transferred to another **strain**[strain-5] of **mouse**[mouse-1], resulting in the latter's production of high levels of antibody molecules it would not normally produce – antibody molecules mimicking the antibody molecules produced by the original **strain**[strain-5],"investigators said in a written statement. (...) after reviewing scientific evidence and performing a **computerized** statistical analysis that showed the false data was not made up of chance errors (...)

(b) Robust-WSD datu-multzoko LA112694-0025 dokumentua desanbiguatuta.

SOUNDS: LATEST WORK IS BOWEN'S MOST HIGH-PROFILE; COMPOSER AND PERFORMER OF NEW MUSIC SPENT YEARS WORKING ON THE FRINGES.
Listening to the lilting **strains**[strain-3] of Gene Bowen's new album "The Vermilion Sea"(...) the Nordic-looking Bowen has a few guitars, a synthesizer and the all-important **computer** – his main composing tool – and piles of records and CDs. (...) Three years ago, Bowen began his work-in-progress, creating the raw material on synthesizers and **computers**. (...) "My interests came through guitar music and songwriting coupled with interest in folk and ethnic music, where **repetition** is always so important. **Repetition** and texture are almost more important than (...)

(c) Robust-WSD datu-multzoko LA063094-0099 dokumentua desanbiguatuta.

2 FIRMS ADOPT LABELS WARNING **COMPUTER** USERS ABOUT DANGER OF **INJURY**. SAFETY GUIDES PROVIDE USERS WITH TIPS.
Compaq **Computer** Corp. said Tuesday that it will put warning labels on **computer** keyboards this fall, directing people to read a safety guide with tips to avoid hand and wrist **injuries**.(...) **Injuries** can range from simple soreness to a tissue swelling that harms nerves in the wrist, a condition known as carpal tunnel syndrome. (...) Compaq said Tuesday that there is still no scientifically established link between keyboard design and **injuries**. But it cited growing evidence, chiefly in news accounts, that typing with hands in awkward positions or for a long time can be harmful. (...) Microsoft has built a healthy side business in **computer** accessories, such as an ergonomic **mouse**[mouse-4] control. (...)

(d) Robust-WSD datu-multzoko LA081794-0225 dokumentua desanbiguatuta.

1.6 irudia – Kontsulta eta dokumentu desanbiguatuak (*mouse* eta *strain* hitzak bakarrik).

rekaketa eginez gero, bietan errepikatzen diren hitz kopuru erlatiboa kopuru osoarekiko txikiagoa izango da, kontsultaren hedapeneko hitzak dokumentuan ez direlako agertuko, eta alderantziz. Ondorioz, IB sistemak dokumentu hori beheragoko postu batean itzuliko du. Baina, 1.6d dokumentuaren hedapenak kontsulta horren hedapenaren antz handia izango du, hitz berdinen kopurua handiagoa izango da, eta dokumentua goragoko postuan jarriko du.

Hemen sinonimia (ahaidetasuna barne) eta polisemia bi fenomeno desberdin bezala aurkeztu baditugu ere, normalean biak nahasirik agertzen dira. Arestian 1.2b adibidea sinonimiaren adibidetzat jarri dugu, *cook* eta *bake* hitzaren artean dagoen hiperonimia-hiponimia erlazioa dela eta. Baina, *cook* hitza polisemikoa da, aditz moduan adiera bat baino gehiago ditu eta; adierarik erabiliena *janaria prestatze*arena bada ere (adibideko kasua ere hala da), *faltsutze*a ere esan nahi du, WordNeten arabera.

Laburbilduz, adibideen bidez ohar gaitezke kontsulta eta dokumentuak hitz-segida moduan ikusi beharrean, hitz horien esanahia ere kontuan hartuz gero berreskurapenean arrakasta lortzeko aukerak handiagoak izango direla. Horregatik, HP teknikak, edo zehatzago esanda, lexiko-semantikako teknikak IBko ad hoc atazan txertatzen saiatuko gara. Tesi-lan honetan **hitzen adiera-desanbiguazioa eta ahaidetasun semantikoa erabiliko ditugu kontsulta eta dokumentuak hobeto *ulertzeko*, eta hedapena erabiliz informazio hori IB sisteman txertatzeko**. Horrela, lexiko-semantikaren bidez kontsulta- eta dokumentu-hedapena eginez, hitz berriak lortu eta hitz horiek IB sistemaren baitan txertatu eta berreskurapen-prozesutik lortzen diren dokumentu adierazgarri gehiago goragoko posizioetan berreskuratzea da gure helburu nagusia. Are gehiago, erabiliko dugun hedapen-teknikak kontsulta eta dokumentuak itzultzeko balio duenez, hedapen-teknika hori erabiliz hizkuntza arteko berreskurapenean hobekuntzak lortzerik ba ote dagoen ere aztertu nahi dugu.

1.4 Aurrekariak IXA taldean

Tesi-lan hau IXA taldearen jardunean kokatzen da. Euskal Herriko Unibertsitateko IXA ikerketa-taldeak hogei urte baino gehiago daramatza HParen arloan lanean. Talde honen xede nagusia euskararen gaineko ikerketa aplikatua den arren, beste hizkuntzen inguruko ikerketa eta produktuen garapena ere jorratzen du.

Talde honetan IBaren arloa asko landu ez bada ere oraindik, EusBila

izeneko euskararako bilaketa-zerbitzu bat garatu da (Leturia *et al.*, 2007). Lexiko-semantika arloari dagokionez, hainbat baliabide eta sistema garatu, eta beste hainbat tesi eta lan argitaratu dira, bai HADaren arloan (Agirre, 1999; Martinez, 2004; Lopez de Lacalle, 2009), eta baita ahaidetasun semantikoaren inguruan ere (Agirre *et al.*, 2009c). Alor honen baitan, euskararako zenbait baliabide ere garatu dira: EuSemcor (semantikoki etiketatutako euskarazko corpusa) eta Euskal WordNet (Pociello, 2008). Aipatu berri ditugun lan eta tresna batzuez baliatu gara tesi-lan honetan.

Hizkuntzari dagokionean, tesi honetan, batez ere, ingeleserako ebaluatu ditugu gure teknikak, ebaluazioa euskaraz egiteko datu-multzorik ez zegoelako. Salbuespenetako bat ResPubliQA atazako euskara-ingelesea ariketa izan da. Hala ere, tesi-txosten honetan aurkeztuko ditugun teknikak hizkuntzarekiko independenteak dira, eta hortaz, edozein hizkuntzatara aplika daitezke baliabide nahiko izanez gero hizkuntza horretarako.

1.5 Ikerketa-galderak

Tesi-lan honen ardatza honako galdera hau izango da:

> **Kontsulten eta dokumentuen hedapenerako semantika lexikala erabiliz hobetzen al da IB sistemen eraginkortasuna ad hoc atazetan?**

Labur esanda, kontsulta eta dokumentuen hedapena eginez, berauek aberastu nahi ditugu parekatze-arazoari aurre egin nahian. Hedapena egiteko semantika lexikaleko bi teknikaz baliatuko gara: hitzen adiera-desanbiguazioa eta ahaidetasun semantikoa. Alde batetik, teknika hauetako bakoitzerako hedapen-prozesu bat proposatuko dugu, non kontsulta eta dokumentuetako hitzen sinonimo eta bestelako ahaidetasuna duten hitzak lortuko ditugun. Bestetik, hedapenetik lortutako hitz horiek, kontsulta eta dokumentuetako jatorrizko hitzekin batera, IB sistemaren prozesuan txertatu eta ustiatzeko modu bat azalduko dugu kasu bakoitzerako. Hau dena kontsulta eta dokumentu adierazgarrien arteko bat etortzea handiagoa izateko helburuarekin.

Galdera nagusi horri eta ondorengo beste ikerketa-galdera (IG) zehatzago hauei erantzuna bilatzen saiatuz bideratuko dugu ondorengo kapituluetan azalduko dugun lana.

– **IG 1 – Hitzen adiera-desanbiguazioan eta ezagutza-base lexikal bateko sinonimoetan oinarritutako hedapena eginez hobetzen al da IB sistemaren eraginkortasuna?**

- 1.1 - Hedapen-teknika hau egokia al da kontsulta- zein dokumentu-hedapenerako? Bi hauetakoren bat ba al da bestea baino eraginkorragoa?
- 1.2 - Hedapen-teknika honen eraginkortasunean zein faktorek eduki dezakete eragina?
- 1.3 - Hedapen-teknika hau egokia al da kontsulten eta dokumentuen itzulpena egiteko hizkuntza arteko berreskurapenean?

– **IG 2 – Ezagutza-base lexikal bidezko ahaidetasun semantikoan oinarritutako hedapena eginez hobetzen al da IB sistemaren eraginkortasuna?**

- 2.1 - Hedapen-teknika hau eraginkorra al da izaera desberdineko berreskurapen-ereduetarako?
- 2.2 - Hedapen-teknika hau egokia al da kontsulta- zein dokumentu-hedapenerako? Bi hauetakoren bat ba al da bestea baino eraginkorragoa?
- 2.3 - Hedapen-teknika hau *pseudo-relevance feedback* metodoarekin alderatzean, zer ikusten dugu?
- 2.4 - Hedapen-teknika honen eraginkortasunean zein faktorek eduki dezakete eragina?
- 2.5 - Hedapen-teknika hau egokia al da kontsulten eta dokumentuen itzulpena egiteko hizkuntza arteko berreskurapenean?

1.6 Tesi-txostenaren eskema

Honakoa da tesi-txosten honen egitura:

- 1. kapitulua – Tesi-lanaren aurkezpen orokorra:

  Irakurtzen ari zaren kapitulu honetan, lehenik, tesi-lan honetan aztertu dugun gaiaren aurkezpen orokorra egin dugu, tesi-lan hau egitera zerk

bultzatu gaituen eta gure helburu nagusia zein den azalduz. Jarraian, gure ikerketekin erantzuna bilatu nahi izan diegun galderak zeintzuk diren mahaigaineratu dugu. Honen ondoren, tesi-lan honekin lotutako argitalpenak zerrendatu ditugu.

- 2. kapitulua – Artearen egoera:

  Hasteko, IB sistema batean ranking-funtzioak duen garrantzia dela eta, ranking-funtzio baten atzean egon daitezkeen berreskurapen-eredu desberdinak zeintzuk diren eta hauen bilakaera nolakoa izan den ikusiko dugu. Horren ostean, parekatze-arazoa zertan datzan ikusiko dugu, eta horri aurre egiteko irtenbide posibleak eta semantika erabiliz egin izan diren saiakeren berri emango dugu.

- 3. kapitulua – Esperimentazio-ingurunea:

  Tesi-txostenean azalduko ditugun esperimentuak garatzeko esperimentazio-ingurunea deskribatu eta hurrengo kapituluetan kontatuko dugun a ulertzeko hainbat oinarrizko kontzeptu azalduko ditugu. Honela ba, hauek izango dira jorratuko ditugun gaiak: informazioaren berreskurapeneko oinarrizko ataza nola egikaritzen den eta horretarako erabil daitezkeen algoritmoak eta tresnak zeintzuk diren, gure esperimentuetan erabilitako hizkuntzaren prozesamenduarekin zerikusia duten hainbat teknika eta baliabide deskribatu, esperimentuak egiteko erabilitako datu-multzoak aurkeztu eta ebaluazioaren inguruko hainbat xehetasun eman.

- 4. kapitulua – Adiera-desanbiguazioa eta hizkuntza-ereduetan oinarritutako IBa:

  Hitzen adiera-desanbiguazioa erabiliz IB sistemaren eraginkortasuna hobetzeko helburuarekin egindako esperimentuak azalduko ditugu. Horretarako, hitzen adiera-desanbiguazioa eta ezagutza-base lexikal bat (WordNet) baliatuz, kontsulta eta dokumentuei sinonimoak gehituz hedapena egitea proposatzen dugu. Ataza elebakarreko (ingelesa) eta hizkuntza arteko atazako (gaztelania-ingelesa) esperimentuak egin ditugu; esperimentu hauek *Robust WSD Task @ CLEF 2008* atazan parte hartzeko egin genituen (Agirre *et al.*, 2009a).

- 5. kapitulua – Ahaidetasuna eta IB probabilistikoa:

  Ahaidetasun semantikoa erabiliz eredu probabilistikoan oinarritutako IB sistemaren eraginkortasuna hobetzeko helburuarekin egindako esperimentuak azalduko ditugu. Horretarako, ezagutza-base lexikal bat (WordNet) oinarritzat duen grafo-algoritmo baten bidez, dokumentuak hedatuko ditugu berauekin erlazionatutako hitzak gehituz.

- 6. kapitulua – Ahaidetasuna eta hizkuntza-ereduetan oinarritutako IBa:

  Ahaidetasun semantikoa erabiliz hizkuntza-ereduetan oinarritutako IB sistemaren eraginkortasuna hobetzeko helburuarekin egindako esperimentuak azalduko ditugu. Horretarako, aurreko kapituluko esperimentuetan erabilitako teknika jarraituz, kontsulta eta dokumentuak hedatuko ditugu.

- 7. kapitulua – Ondorioak eta etorkizuneko lanak:

  Alde batetik, aurreko kapituluetan egindako esperimentuetatik ateratako ondorioak eta tesi-lan honen ekarpenak zeintzuk izan diren laburbilduko ditugu. Bestetik, lan honen inguruan etorkizunean zein bide har daitezkeen zehaztuko dugu.

Kapitulu hauez gain, glosategia eta beste honako eranskin hauek aurki daitezke tesi-txosten honetan:

- A eranskina – *Stopword*en zerrenda.

- B eranskina – Adiera-desanbiguazio bidezko hedapeneko fitxategiak:

  4.2 ataleko adibideei dagozkien XML fitxategiak aurki daitezke eranskin honetan.

1.7 Tesi honen garapenetik atera diren argitalpenak

Tesi-lan honek hainbat artikulu argitaratzeko bidea eman digu. Hona hemen argitalpen horien zerrenda, kapituluen arabera sailkatuta[6]:

[6]Oharra: Argitalpen hauetan autoreen ordena alfabetikoa da, 6. kapituluari dagokionean izan ezik.

- 4. kapitulua – Adiera-desanbiguazioa eta hizkuntza-ereduetan oinarritutako IBa:

  - Agirre E., Otegi A., eta Rigau G. **IXA at CLEF 2008 Robust-WSD Task: Using Word Sense Disambiguation for (Cross Lingual) Information Retrieval**. *Evaluating Systems for Multilingual and Multimodal Information Access, CLEF 2008*, Lecture Notes in Computer Science, 5706 lib., 118–125. Springer, ISBN 978-3-642-04446-5. 2009.

- 5. kapitulua – Ahaidetasuna eta IB probabilistikoa:

  - Agirre E., Otegi A., eta Zaragoza H. **Using semantic relatedness and word sense disambiguation for (CL)IR**. *Multilingual Information Access Evaluation I - Text Retrieval Experiments, CLEF 2009*, Lecture Notes in Computer Science, 6241 lib., 166–173. Springer, ISBN 978-3-642-15753-0. 2010.
  - Agirre E., Ansa O., Arregi X., Lopez de Lacalle M., Otegi A., Saralegi X., eta Zaragoza H. **Elhuyar-IXA: Semantic Relatedness and Cross-Lingual Passage Retrieval**. *Multilingual Information Access Evaluation I - Text Retrieval Experiments, CLEF 2009*, Lecture Notes in Computer Science, 6241 lib., 273–280. Springer, ISBN 978-3-642-15753-0. 2010.
  - Agirre E., Ansa O., Arregi X., Lopez de Lacalle M., Otegi A., eta Saralegi X. **Document Expansion for Cross-Lingual Passage Retrieval**. *Proceedings of CLEF 2010 Workshop on Multiple Language Question Answering (MLQA'10)*, ISBN 978-88-904810-0-0. 2010.
  - Agirre E., Arregi X., eta Otegi A. **Document expansion based on WordNet for robust IR**. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, 9–17, Association for Computational Linguistics. 2010.

- 6. kapitulua – Ahaidetasuna eta hizkuntza-ereduetan oinarritutako IBa:

  - Otegi A., Arregi X., eta Agirre E. **Query Expansion for IR using Knowledge-Based Relatedness**. *Proceedings of the 5th*

International Joint Conference on Natural Language Processing, 1467–1471, ISBN 978-974-466-564-5. 2011.

Beste hauek ere, kapitulu zehatz batekin lotura ez badute ere, tesiarekin zerikusia dute:

– Agirre E., Magnini B., Lopez de Lacalle O., Otegi A., Rigau G., eta Vossen P. **SemEval-2007 Task 01: Evaluating WSD on Cross-Language Information Retrieval**. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 1–6. 2007.

– Agirre E., Di Nunzio G.M., Mandl T., eta Otegi A. **CLEF 2009 Ad Hoc Track Overview: Robust-WSD Task**. *Multilingual Information Access Evaluation I - Text Retrieval Experiments, CLEF 2009*, Lecture Notes in Computer Science, 6241 lib., 36–49, Springer, ISBN 978-3-642-15753-0. 2010.

2

Artearen egoera

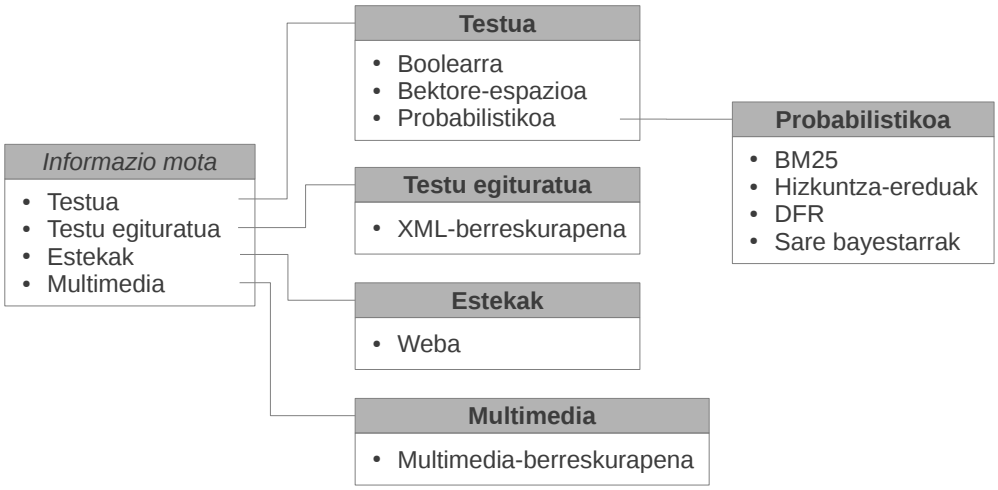
IB sistema baten ezaugarri behinena berreskurapen-eredua da. Hori dela eta, kapitulu honetako lehenengo atalean berreskurapen-eredu desberdinak azalduko ditugu, arreta berezia jarriz ranking-funtzioetan, berebiziko garrantzia baitute. Bigarren atalean, berreskurapen-eredu soilak erabiliz sortu ohi den parekatze-arazoa zertan datzan ikusiko dugu. Azkenik, horri aurre egiteko irtenbide posibleak eta semantika erabiliz egin izan diren saiakeren berri emango dugu.

2.1 Berreskurapen-ereduen bilakaera

IBaren hastapena 40-50eko hamarkadetan kokatu ohi da. Lan zientifikoen argitalpen kopuruaren gorakadarekin eta ordenagailuen erabilera areagotzearekin bat, dokumentuen berreskurapena egiteko interesa piztu zen. Hori bai, egile, titulu eta gako-hitzetan oinarritutako bilaketak izango ziren garai hartako bilaketak; dokumentuen eduki osoa aztertuz egindako bilaketak geroago etorri ziren (Manning *et al.*, 2009).

Orduz geroztik, hainbat eta hainbat berreskurapen-eredu garatu eta erabili izan dira. Erabiltzaileak egindako kontsulta bat emanik, berreskurapen-eredu batek aurreikusi beharko luke zein dokumentu izango diren adierazgarriak erabiltzailearentzat, horretarako arrazoizko biltegiratze-espazioa erabiliz eta erantzuna arrazoizko denboran emanez.

2.1 irudian IB ereduen sailkapen bat ikus daiteke, berreskuratzen den dokumentu motaren arabera antolatua: testua, estekak eta multimedia. Orain



2.1 irudia – IB ereduen sailkapena.

arte gehien landu eta erabilitakoa testuaren berreskurapena izan denez, mota horretako ereduen bilakaera kronologikoa zein izan den ikusiko dugu jarraian. Horrela bada, eredu boolearrak, bektore-espazioaren ereduak eta eredu probabilistikoak azalduko ditugu, bereziki azken horiek jorratuko ditugularik, mota horretako ereduak izan baitira lan honetan erabili ditugunak. (Baeza-Yates eta Ribeiro-Neto, 2011) liburuan eredu nagusienen eta hemen aipatu gabe utziko ditugun beste hainbat ereduren kontakizun bibliografiko osatuago bat aurki daiteke.

Eredu boolearrak

Lehenengo bilatzaileek erabiltzen zuten berreskurapen-eredua boolearra zen. 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. Eta kontsultako termino horiek dokumentuetan egote edo ez egotean oinarritzen da berreskurapen-eredua. Alegia, kontsultako adierazpen logikoa egiazko egiten duten dokumentuak izango dira berreskuratuko direnak (adierazgarriak), eta gainontzeko dokumentuak baztertu egingo dira (ez-adierazgarriak). Hortaz, eredu honetan dokumentuak bi multzotan sailkatzen dira —adierazgarriak eta ez-adierazgarriak—, eta, beraz, ez da

dokumentuen mailakatzerik edo rankingik egiten.

Abantailaren bat edo beste badu, eta bat aipatzearren, ereduaren sinpletasuna azpimarratuko genuke. Desabantailak ere baditu, ordea. Desabantailarik handiena dokumentuen rankinga ez egitea da. Bestalde, kontsultako espresio osoa bete behar du dokumentu batek adierazgarria izateko, hots, parekatze partzialik ez da onartzen. Horretaz gain, dokumentu batean termino bat agertzen den edo ez bakarrik hartzen da kontuan, eta ez zenbat aldiz agertzen den termino hori. Gainera, erabiltzaileak duen informazio-behar hori eragile logikoen bidez kontsulta batean adieraztea zaila suerta liteke batzuetan (Baeza-Yates eta Ribeiro-Neto, 2011).

Gaur egun oraindik sistema batzuek eredu hau erabiltzen jarraitzen badute ere, 1960ko hamarkadatik aurrera indar handia hartu zuten aipatu berri ditugun eragozpen horiek izango ez zituzten ereduek. Jarraian ikusiko ditugun ereduak estatistikoak dira, dokumentuetako terminoen agerpen kopuruak hartzen baitituzte kontuan rankingak egiteko. Kontsulta-adierazpenak automatikoki sortzen dira; erabiltzaileak informazio-beharra hizkuntza arruntean idatzi ahal izango du edo termino gutxi batzuk idatzi, eragile logikoak erabili beharrik izan gabe.

Bektore-espazioaren ereduak

Luhn (1957) izan zen informazio-bilaketaren prozesuari ikuspuntu estatistiko bat eman behar zitzaiola proposatzen lehena. Luhn-ek zioen zuzenean dokumentu-bilduman bilatu beharrean, hobe zela dokumentu bakoitza eta kontsulta eduki-identifikadore batzuen bitartez errepresentatzea, eta bi errepresentazioen arteko antzekotasun-maila erabiltzea berreskurapenerako irizpide moduan. Hots, eredu honen arabera, errepresentazioetan agertzen diren elementuak zenbat eta gehiago izan eta hauen banaketa zenbat eta berdinagoa izan, orduan eta handiagoa izango da bi errepresentazioek informazio bera errepresentatzeko probabilitatea. Hortaz, Luhn bera izan zen gakohitzetan oinarritutako indizeen bitartez bilaketak egitea proposatu zuen lehena.

Antzekotasun-irizpide horretan oinarrituz, Salton-ek (1971b) 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 desberdin bakoitzeko bektoreko elementu bat izango dugu, eta elementu bakoitzak pisu bat izango du.

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

Termino-haztapena

Esan dugun moduan, bektoreko elementu bakoitzak pisu bat izango du. Pisu hau balio bitarra izan badaiteke ere —eredu boolearrean bezala, termino baten agerpena 1eko baten bidez adieraziko litzateke kasu honetan—, eredu honek aukera ematen du termino bakoitzari beste nolabaiteko pisuak esleitzeko, estatistiketan oinarritutakoak, kasu. Horrelakoetan erabili ohi diren bi balio ezagunenak *tf* eta *idf* dira. *tf* ingelesezko *term frequency*tik dator, eta dokumentu batean termino batek duen agerpen kopurua adierazten du. *idf* (*inverse document frequency*) edo dokumentu-maiztasunaren alderantzikatua deituko diogun balioak 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 oso ohikoa den terminoak (Robertson, 2004). Termino batentzat *tf* altua duten dokumentuak termino hori duten kontsulta batekiko adierazgarriak izango direla pentsa liteke. Baina, gainera, bilduman ohikoak ez diren terminoak baztertzaileagoak dira eta informazio gehiago ematen dute asko agertzen diren terminoek baino (Jones, 1972). Horregatik, termino bati pisua esleitzeko oso erabilia da *tf-idf* neurketa, aipatu berri ditugun *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 eredu honetan bezalaxe, jarraian ikusiko ditugun beste IB eredu gehienetan ere.

Eredu probabilistikoak

Bektore-espazioaren ereduak dokumentuen rankinga egiten badu ere, hori egiten lehenak Maron eta Kuhns (1960) izan zirela esan ohi da, IBaren

arloan erabat berria izango zen ikuspuntu probabilistiko batean oinarrituz. IBaren arlora adierazgarritasunaren kontzeptua beraiek ekarri zuten, honakoa adieraziz: IB sistema batek dokumentu bakoitzari kontsultarekiko duen adierazgarritasun-probabilitate bat esleitu beharko lioke eta probabilitate horien arabera ordenatu dokumentuak. Ideia honetan oinarritzen da gaur egungo eredu probabilistiko guztien jatorrian dagoen *probability ranking principle* deituriko printzipioa (Robertson, 1977):

> "If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data." (W.S. Cooper)

Autorearen arabera, dokumentuek kontsultarekiko adierazgarriak izateko duten probabilitatearen arabera ordenatzean lortzen da IB sistema baten errendimendurik hoberena. Hortaz, eredu probabilistiko guztien helburua Q kontsultaren arabera D dokumentuak $P(R = 1|Q, D)$ probabilitatearen arabera ordenatzea izango da. Printzipio honetan adierazgarritasuna (ingelesezko *relevance*tik R moduan adierazi ohi dena) aldagai bitartzat hartzen da, eta, beraz, $R = 1$ izango da D dokumentua adierazgarria denean Q kontsultarentzako, eta 0 bestela. Gainera, dokumentuak bata bestearekiko independentetzat hartzen dira, dokumentu baten adierazgarritasuna bakarrik dokumentu horren araberakoa dela pentsatuz.

Printzipio honetan ez da zehazten adierazgarritasunaren estimazioa nola egin behar den. Hori dela eta, printzipio honetan oinarriturik hainbat eta hainbat eredu probabilistiko garatu dira, eta eredu horien arteko alderik handienak probabilitate hori kalkulatzeko moduan daude. Printzipio honi jarraituz sortu zen lehen eredua *binary independence model* izenez ezagutzen dena da (Robertson eta Jones, 1976; van Rijsbergen, 1979). Eredu honek sinplifikazio batzuk onartzen ditu $P(R|Q, D)$ kalkulatzea posible izateko. Alde batetik, proposatzen duten eredua bitarra da —eredu boolearra bezala— eta dokumentuak eta kontsultak bektore bitar moduan adierazten dira —dokumentu baten bektoreko elementuak 1 balioa izango du elementu horri dagokion terminoa dokumentuan baldin badago eta 0 bestela; berdin

kontsultaren bektorean ere—. Bestetik, dokumentu batean terminoak bata bestearekiko independenteak direla onartzen da. Gainera, eredu honek onartzen du kontsulta bat emanik dokumentu adierazgarrien multzoa eta ez-adierazgarriena ezagunak direla. Hemendik RSJ (Robertson-Sparck Jones) izenez aski ezaguna den pisua edo formula lortu zuten. Hitz gutxitan esanda, teknika estatistikoen bidez adierazgarritasun-informazioaz baliatzen dira bilaketa-terminoei pisu bat emateko[1].

Esan dugun bezala, kalkulu horiek egiteko aurretiaz adierazgarritasun-informazioa beharrezkoa da. Jakina den moduan, gehienetan informazio hori ez da eskura izaten bilaketa egin aurretik. Kasu hauetan, bildumako dokumentu guztiak ez-adierazgarriak direla suposa dezakegu —ohikoena dokumentu asko izan eta horietatik adierazgarriak gutxi izatea da; beraz, proportzioak ikusita, ez da hain zentzugabekoa hori suposatzea—, eta orduan formula horrek hartzen duen balioa *idf* balioaren oso antzekoa da. Hortaz, eredu honek horrelako egoera batean —egoera erreala kasurik gehienetan hori da— dokumentu bati esleitzen dion pisua ez da oso ona izango, eta dokumentuen rankinga *tf-idf* neurri sinplearekin egindako rankinga baino okerragoa izango litzateke (Croft *et al.*, 2009).

Orain arte kontatutakoa eredu probabilistiko guztien muina bada ere, hortik hainbat eredu eta algoritmo desberdin garatu dira; besteak beste, BM25, honen aldaera bat den BM25F, hizkuntza-ereduak, DFR eta sare bayestarrak. Ondoren, horiexen berri emango dugu laburki.

BM25 izenez ezagutzen den ranking-funtzioak[2] dokumentuan terminoaren agerpen kopurua (tf) eta dokumentuaren luzera hartzen ditu kontuan (Robertson eta Walker, 1994). IBaren arloan arrakasta handia izaten ari da algoritmo hau, datu-multzo eta ataza desberdinetan oso emaitza onak lortu baititu (hainbat esperimenturen emaitzak eta ondorioak aurki daitezke (Jones *et al.*, 2000) lanean). Gure esperimentu batzuetan hau erabili dugunez, 3.2.1 atalean sakonago ikusiko dugu ranking-funtzio hau.

BM25F algoritmoa etorri zen ondoren, BM25 algoritmoaren egokitzapen sinple bat dena. Hau egokia da eremu bat baino gehiagoko dokumentuetan (HTML dokumentuak, esaterako) eremu batzuek beste batzuek baino garrantzi handiagoa dutela uste denerako (Robertson *et al.*, 2004).

[1] Hemen ez dugu matematikoki nola garatu zuten jarriko; informazio hori, besteak beste, (Robertson eta Jones, 1976) eta (Robertson eta Zaragoza, 2009) lanetan aurki daiteke.

[2] Okapi BM25 izenarekin ere ezagutzen da, Okapi izeneko testuen berreskurapen-sistemarako inplementatu baitzen lehenengo aldiz.

Badago orain arte ikusi ditugun eredu probabilistiko klasikoetatik urruntzen den beste eredu probabilistiko bat, hain zuzen ere, hizkuntza-ereduetan oinarritzen den eredua. Hizkuntza-eredua (*language model*) hitz-segidei probabilitate-banaketak esleitzen dizkien eredu probabilistiko bat da. Hizkuntzaren prozesamenduan asko erabiltzen da, besteak beste, hizketaren ezagutzarako eta itzulpen automatikorako. Ponte eta Croft (1998) izan ziren lehenak hizkuntza-ereduak IB sistema batean txertatzea proposatu zutenak. Honako hau da eredu honen atzealdean dagoen ideia: dokumentu baten hizkuntza-eredutik kontsulta sortu baldin badaiteke, dokumentu hori kontsulta horrentzat egokia izango da, hau gertatuko baita kontsultako terminoak dokumentuan agertzen badira. Eredu probabilistiko klasikoek $P(R = 1|Q, D)$ probabilitatearen araberako rankingak egiten dituzten bitartean, hizkuntza-ereduetan oinarritutako oinarrizko ereduak D dokumentu bakoitzaren Θ_D hizkuntza-eredua sortu, eta $\Theta_D$tik kontsulta sortzeko dagoen probabilitatearen arabera ordenatuko ditu dokumentuak —$P(Q \mid \Theta_D)$ probabilitatearen arabera—. Eredu konkretu hau *query likelihood* (QL) edo kontsulta-egiantza izenez ezagutzen da, eta 3.2.2 atalean ikusiko dugu zehatzago.

Hori bada ere IBrako hizkuntza-ereduen hurbilpenik oinarrizkoena, badira beste aukera batzuk ere. Esaterako, kontsultaren Θ_Q hizkuntza-eredutik dokumentua sortzeko dagoen probabilitatea erabil daiteke. Eredu honi *document likelihood* eredua deitzen zaio. Baina, ez da hain erabilia, kontsultan oinarritutako hizkuntza-ereduaren estimazioa egiteko testu gutxiago izango baitugu —ohikoena kontsulta oso laburrak izatea da—, eta segur aski, ez da Θ_D bezain ona izango. Baina, Θ_Q hizkuntza-ereduan, Θ_D hizkuntza-ereduan ez bezala, oso erraz txerta daiteke *relevance feedback* edo adierazgarritasun-feedback delakoa: dokumentu adierazgarrietatik hartutako terminoekin hedatu kontsulta eta Θ_Q berriaren estimazioa egin. Lavrenko eta Croft-ek (2001) proposatutako adierazgarritasun-eredua *document likelihood* eredu honetan oinarritzen da, eta oso emaitza sendoak lortzen ditu.

Θ_Q edo Θ_D bietako bat bakarrik erabili beharrean, biak kalkulatu eta bien arteko alderaketa bat ere egin dezakegu dokumentu adierazgarriak berreskuratzeko; adibidez, bi probabilitate-banaketen arteko diferentzia kalkulatzeko erabili ohi den *Kullback-Leibler* (KL) dibergentzia erabil daiteke bi hizkuntza-ereduak konparatzeko (Zhai eta Lafferty, 2001b).

Dagoeneko ikusi dugu IB sistema batean hizkuntza-ereduak erabiltzeko hainbat hurbilpen daudela. Baina hurbilpen horietako bakoitzean ere, beste hainbat aldaera garatu dira. Esaterako, QLaren inplementazioan Θ_D

hizkuntza-eredua nolakoa izango den erabaki behar da (multinomiala, Bernoulliren eredua...). Gainera, garrantzi berezia du $\Theta_D$ren estimazioan leuntze-prozesuak (*smoothing*), eta hainbat leuntze-teknika erabili ohi dira (Dirichlet, Jelinek-Mercer...). Aldaera hauek guztiak eta gehiago aztertu zituen Zhai-ek (2008).

Hizkuntza-ereduak bezalaxe, *divergence from randomness* (DFR) izendatzen diren ereduak ere dokumentuetako terminoen banaketa estatistikoan oinarritzen dira. Honako hau da DFR ereduen atzealdean dagoen ideia: termino baten dokumentu-maiztasunaren eta bilduma-maiztasunaren artean dagoen dibergentzia zenbat eta handiagoa izan, orduan eta ekarpen handiagoa egiten dio termino horrek dokumentu horri. Beste modu batean esanda, termino baten ausazko banaketa eta unean duen banaketaren artean dagoen diferentziaren araberakoa izango da termino horren pisua (Amati eta van Rijsbergen, 2002).

Turtle eta Croft-ek (1991) sare bayestarrean oinarritzen zen inferentzia-sarearen eredua proposatu zuten. Inferentzia-sareen bidez oso erraz inplementa daitezke hainbat eragile probabilistikoz osatutako kontsulta egituratuak[3] (Greiff *et al.*, 1999). Baina, eredu hauek badute hain positiboa ez den alderdi bat: terminoen probabilitateak *tf-idf* pisu soilarekin estimatzen dira. Hau dena kontuan izanik, Metzler eta Croft-ek (2004) inferentzia-sarea eta hizkuntza-ereduak konbinatzea proposatu zuten. Horrela, kontsulta egituratuak onartzen zituen eta terminoen pisuak hizkuntza-ereduetan oinarritutako probabilitateen bidez estimatu ahal izango diren eredu bat izango dugu.

Indri izeneko IB sistema azken eredu berritzaile honetan oinarritzen da. Guk gure esperimentu batzuetarako sistema hori erabili dugu eta 3.3.1 atalean emango dugu sistema honen berri zehatzagoa.

Hainbeste berreskurapen-eredu izanda, hainbat esperimentu egin izan dira probatzeko ea hauetako ereduren bat beste bat baino hobea ote den, baina desadostasunak daude (Baeza-Yates eta Ribeiro-Neto, 2011). Zhai-ek (2008) esaten duenez, *query likelihood* berreskurapen-funtzioak, *Dirichlet* leuntze-teknika aplikatuz, BM25 funtzioaren pareko errendimendua lortzen du. Voorhees-en (1999) arabera, berreskurapen-eredu oso desberdinak erabiltzen dituzten sistemek antzeko eraginkortasuna lortzen dute berreskurapenean, beti ere terminoen haztapen egokia egiten bada. Izan ere, terminohaztapenak funtsezko ondorioak ditu berreskurapenaren kalitatean, emai-

[3]Kontsulta egituratu bat ez da termino-segida soil bat, baizik eta terminoak multzokatu daitezke, edo erlazioak jarri, pisu desberdinak esleitu terminoei, etab.

tzarik onenak *tf*, *idf* eta dokumentuen luzera konbinatuz lortzen direlarik. Hiemstra-k (2009), berriz, dio eredu batzuk egokiagoak direla IBko aplikazio jakin batzuetarako, eta beste eredu batzuk beste aplikazio batzuetarako.

2.2 Parekatze-arazoa

Aurreko atalean ikusi ditugun berreskurapen-ereduak elkarren artean guztiz desberdinak dira, baina guztiak bat datoz honako honetan: kontsulta eta dokumentuak komunean dituzten hitz kopuruaren araberakoa da dokumentu baten adierazgarritasuna. Hitzak komunean izan behar dituztela esatean, hitz horiek lexikografikoki berdinak izan behar dutela esaten ari gara. Baina sarreran ikusi dugun moduan, bi termino lexikografikoki desberdinak izanagatik semantikoki antzekoak izan daitezke, eta lagungarriak izan dokumentu baten adierazgarritasuna zehazteko. Eta alderantziz, lexikografikoki berdinak diren hitzak ez dira beti dokumentu adierazgarri baten adierazle izango. Hauxe da atal honetan jorratuko dugun parekatze-arazoaren oinarria.

IBrako indizeak erabiltzen hasi zirenarekin bat ohartu ziren berreskurapena ondo egitea ez zela lan samurra izango. Swanson-ek (1988) dioenez, O'Connor (1961) 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, kontsultan hitz hori erabiliz gero, litekeena baitzen 11 dokumentu horiek ez lortzea.

Kasu horretan eta beste hainbatetan kontsulta eta dokumentuen artean egon zitekeen hutsune horretaz jabetuz, Blair eta Maron-ek (1985) zera adierazi zuten: erabiltzaile batentzat oso zaila gerta liteke dokumentu adierazgarri guztietan (edo gehienetan) agertu eta beste dokumentuetan (ez-adierazgarrietan) agertzen ez diren hitzak edo hitz-segidak zein diren aurreikustea.

Lan berriagoetara etorriz, kontsultako eta dokumentu adierazgarrietako terminoen arteko parekatze-eza gertatzen dela ziurtatzeko, Muller eta Gurevych-ek (2009) TRECeko datu-multzo baten gainean (datu-multzo honetako kontsulten batez besteko luzera 2,44koa zen), kontsultetako eta dokumentu adierazgarrietako terminoen arteko gainjartzea zenbatekoa zen kalkulatu zuten, eta hauek izan ziren emaitzak: dokumentu adierazgarrien % 35,5 dokumentuetan kontsultako termino guztiak agertzen ziren, eta % 86,5 do-

kumentuetan gutxienez kontsultako termino bat agertzen zen. Honenbestez, bilduma horretako % 13,5 dokumentu adierazgarrietan ez da kontsultako terminorik agertzen, eta, hortaz, kontsultako terminoen parekatze soil baten bidez ezingo dira dokumentu horiek berreskuratu.

2.3 Parekatze-arazoari aurre egiten lexiko-semantika erabiliz

HPa erabiliz hainbat irtenbide daude lehen aipatu dugun arazoaren aurrean. Sinpleenak eta emaitza onak ematen dituzten teknikak *stemming* deiturikoa eta lematizatzea dira. Bi teknika hauen helburua hitzen erroak edo lemak lortzea da —bi teknikak oso antzekoak dira, baina jarraitzen duten prozesua desberdina da eta 3.4.5 atalean azalduko ditugu. Hori eginez kontsulta eta dokumentuetako hitzekin, hauen artean parekatze gehiago egoteko aukera egongo da. Ingeleserako *stemmer*ik erabiliena Porter-en (1980) algoritmoa da. Ingeleseko *stemmer*a erabiltzearen alde on eta txarrak hainbat lanetan aurki daitezke: (Salton, 1989), (Harman, 1991), (Hull, 1996) eta (Hollink *et al.*, 2004). Leturia *et al.*-ek (2008) euskarazko bilatzaile batean lematizazioak duen garrantzia azpimarratzen dute. Lematizazioa egiterik izan ez eta horren ordez kontsultaren hedapen morfologikoa eginez lortzen diren emaitzak nolakoak diren aztertzen dute.

Teknika horietaz gain, badaude beste teknika konplexuago batzuk nahiko erabiliak direnak parekatze-arazoari aurre egiteko. Horietako bat hedapena egitea da. Ikusiko dugun moduan, hainbat modutara egin daiteke hedapena. Aukeretako bat HPko tresna orokor edo lexiko-semantikako teknika zehatzagoren baten bidez hedapena egitea baldin bada ere, ez dago zertan horrela izanik, eta estatistika neurri batzuetan oinarrituta egin daiteke hedapena. Jarraian, hedapen mota desberdinak azalduko ditugu.

Hedapena egiteko modu bat kontsulta-hedapena (ingelesez *query expansion*, QE) deritzona da. Kontsulta-hedapeneko metodoek erabiltzaileak egindako kontsultako terminoak aztertu eta hauekin erlazionatutako terminoak gehitzen dizkiote hasierako kontsultari (Voorhees, 1994). Metodo hauek bi multzotan banatu ohi dira: metodo lokalak eta globalak (Xu eta Croft, 1996).

Metodo lokalak hasierako kontsultarekin parekatzen diren dokumentuetan oinarritzen dira kontsulta berri bat sortzeko (Manning *et al.*, 2009). Sasiadierazgarritasun-feedbacka da, ingelesez *blind feedback* edo *pseudo-re-*

levance feedback (PRF) moduan ezaguna dena, mota honetako metodorik erabiliena (Voorhees, 1999). 3.2.3 atalean xehetasun gehiagorekin azalduko badugu ere, labur esanda, metodo honek lehenengo berreskurapen-saiakerako lehen k dokumentuak adierazgarritzat hartuko ditu —eta batzuetan besteak ez-adierazgarritzat—, eta k dokumentu horietan oinarrituko da termino berriak lortzeko. Rocchio-k (1971) adierazgarritasun-feedbackerako algoritmoa aurkeztu zuenetik, hainbat aldaera garatu dira. *TREC 2008 Relevance Feedback Track* izeneko atazako emaitzek baieztatu zuten adierazgarritasun-feedback metodoaren bidez hainbat berreskurapen-eredutan hobekuntzak lortzen direla, baina, erabili beharreko adierazgarritasun-informazioaren kantitatea eta informazio ez-adierazgarria erabili edo ez sistemaren arabera aldatu egiten dela (Buckley eta Sanderson, 2008).

Metodo globalek, aldiz, ez dute hasierako kontsultarekin berreskuratutako dokumentuetan begiratzen, baizik eta terminoen bilduma osoko agerkidetzan (*co-ocurrence*) oinarritutako neurri estatistikoak edo kanpoko ezagutza-iturriren bat erabiltzen dute hedapen-terminoak lortzeko (Manning *et al.*, 2009). Erabiliko diren kanpoko ezagutza-iturri horiek, gehienetan, HPko baliabideak edo teknikak izango dira, eta, adibidez, semantikoki loturaren bat duten terminoak izango dira hedapenean erabiliko direnak.

Are gehiago, kontsultaren hedapena egiten den modu berean, dokumentu-hedapena (*document expansion*, DE) ere egin liteke. Hortaz, aipatu dugun parekatze-arazo horiek konpontzeko ezagutza lexiko-semantikoa kontuan hartuko duten teknika bat baino gehiago jarrai daitezkeela ikusi dugu. Eta teknika gehiago ere badirenez, hona lexiko-semantikako teknikaren baten bitartez lortzen den ezagutza edo informazioa IB sistema batean txertatzeko tekniketako batzuk zerrendatuta:

- Kontsulta- eta dokumentu-hedapena, non kontsultak edo dokumentuak lotura semantikoren bat duten terminoekin hedatuko diren.
- Dokumentuetako hitzekin sortu beharrean indizea, kontzeptuekin sortzea —eta, beraz, kontsulta sortzeko kontzeptuak erabili beharko dira.
- Ezagutza lexiko-semantikoan oinarritutako dokumentuen ranking-funtzioak erabiltzea

Aurreko teknika horietan ezagutza lexiko-semantikoa txertatzen dela esan dugu, eta txertaketa hori hainbat modutara egin daiteke. Aukera batzuk aipatzearren, hitzen adiera-desanbiguaziorako sistema bat (Agirre eta Edmonds, 2006) edota antzekotasun edo ahaidetasun semantikoa neurtzen duen tresnaren (Budanitsky eta Hirst, 2006) bat erabil daiteke.

Tresna edo sistema horiek baliabide lexiko-semantikoren batean oinarri-

tzen dira. Hauek domeinu konkretu batekoak izan daitezke —adibidez medikuntza arloko sistemetan erabili ohi diren UMLS eta MeSH (Medical Subject Headings)[4]— edo ezagutza orokorrekoak —hala nola, WordNet (Fellbaum, 1998), Wikipedia[5], Yago (Kasneci *et al.*, 2009), DBpedia (Auer *et al.*, 2007) edo ConceptNet (Liu eta Singh, 2004).

Esandakoaren arabera, arestian aipatu dugun arazo horri aurre egiteko aukera asko egon daitezkeela ikusi dugu. Azken urteotan saiakera asko egin dira eta HPaz baliatuz egin diren lanik aipagarrienak zeintzuk izan diren ikusiko dugu jarraian; batez ere, HParen baitan kokatzen den lexiko-semantika arloko teknikak (hitzen adiera-desanbiguazioa eta ahaidetasun semantikoa, esaterako) erabili dituzten lanetan jarriko dugu arreta.

Hitzen adiera-desanbiguazioa erabiliz

Hitzen adiera-desanbiguazioa (HAD) erabiliz berreskurapeneko emaitzak hobetzeko saiakeren azalpen zabalak (Sanderson, 2000) eta (Resnik, 2006) lanetan aurki daitezke. Hemen horietako batzuk ikusiko ditugu.

IBaren arloan desanbiguazioarekin egindako probak jasotzen dituen lehenetariko lana (Weiss, 1973) da. Proba hauetan dokumentu-bildumaren errepresentazioa hobetzen saiatu zen HAD sistema bat erabiliz. Lan horretan esaten denez, dokumentu-bildumako hitz anbiguo guztiak ebatziz gero, IB sistema baten eraginkortasuna % 1ean bakarrik hobetuko litzateke; baina, ez da ez esperimentuen xehetasun gehiegi ematen, ezta baieztapen honen argudiorik ere.

IBa eta HADa uztartuz gerora oihartzunik eta eraginik handiena izan duten esperimentuak 90eko hamarkadatik aurrera egindakoak izan dira.

Voorhees-ek (1993) WordNeteko erlazio semantikoetan —hiperonimia eta hiponimia erlazioak bakarrik— oinarritutako desanbiguatzaile bat garatu zuen. Sistema horrekin hainbat bilduma desanbiguatu eta bilduma horiek erabiliz, berreskurapen-sistemaren eraginkortasuna bilduma guztietan jaitsi egiten zela ikusi zuen. Desanbiguazio-sistemaren zehaztasuna (*accuracy*) ez zuen neurtu, baina gainetik begiratuta, ez zela oso zehatza ikusi zuen; eta segur aski hori izan zitekeela emaitza txarrak lortzearen faktorea.

Sussna-k (1993) ere, bide beretik jarraitu zuen, eta WordNeteko erlazioak (guztiak) erabili zituen edozein bi hitz edo adieraren arteko distantzia

[4] http://www.nlm.nih.gov/mesh
[5] http://www.wikipedia.org/

semantikoa kalkulatzeko, eta horrela, desanbiguatzailea sortzeko. Desanbiguatutako hitzen adieratan oinarritutako berreskurapenean emaitza kaxkarrak lortu zituen. Sussna-k bai ebaluatu zuela bere desanbiguazio-sistema; egindako eskuzko ebaluazioan bere sistemak % 56ko zehaztasuna zuela ikusi zuen. Hortaz, zaila da esatea zer dela-eta lortu ote ziren berreskurapeneko emaitza txarrak: desanbiguazioa IB sistema batean txertatzea ez delako egokia edo desanbiguatzailea ez zelako ona.

Wallis-ek (1993) bere IB esperimentuetarako, LDOCE hiztegian oinarritutako desanbiguatzaile bat erabili zuen, eta kontsulta eta dokumentuetako hitz bakoitza hiztegiko definizioan agertzen diren hitzekin ordezkatu zuen. Esperimentu hauetan ere, berreskurapen-sistemaren eraginkortasuna txikiagoa zen desanbiguatzailea erabiliz.

Handik gutxira, Richardson eta Smeaton-ek (1995) antzeko esperimentu bat egin zuten; hauek WordNetetik erauzitako errepresentazio semantikoz ordezkatu zituzten hitzak. Honetan ere, adiera egokiak aukeratzeko desanbiguatzaile bat erabili behar izan zuten. Eta emaitzetan, aurrekoen antzera, ez zuten hobekuntzarik lortu. Baina, aipatutako azken lan hauetan ez zuten erabilitako HAD sistemaren zehaztasuna neurtu, eta hortaz, ezin da ondorio garbirik atera esperimentu hauetatik.

Ondoren etorri zen Smeaton eta Quigley-ren (1996) lana, aurrekoaren oso antzekoa. Baina, eskuz desanbiguatu zituzten kontsultak eta dokumentuak, eta dokumentuak —irudi-oinak ziren— oso motzak ziren. Esperimentu hauetan bai lortu zutela emaitzak hobetzea.

Orain arte aipatu ditugun ia lan guztietan ez zen hobekuntzarik lortu desanbiguatzailea erabilita; baina, lan horietatik ezin izan da ondorio garbirik atera. Jarraian aipatuko ditugun lanetan, ordea, IB bildumetako anbiguotasunaren inguruko hainbat analisi egin zituzten, eta uste bezalako hobekuntzarik zergatik ez zen lortzen argitzeko bidea eman zuten.

Krovetz eta Croft-ek (1992) egindako esperimentuan honako hau ikusi zuten: (i) kontsultako hitzen eta dokumentu adierazgarrietako hitzen arteko adieretan parekatze handiagoa zegoela ez-adierazgarrietan baino, eta (ii) lehenengo postuetan berreskuratutako dokumentuetako hitzen adierak kontsultetako hitzen adierekin bat zetozela gehienetan. Eta hori ondorengo arrazoi hauengatik gerta zitekeela esan zuten:

– Kontsultako hitzen kolokazioaren eragina; alegia, kontsultako hitzek elkar desanbiguatzen dute. Hau argitzeko, hona hemen adibide bat: “banku moneta truke” kontsultarekin berreskuratuko diren dokumentuetan ziurrenik hitz horiek ere agertuko dira eta, ondorioz, dokumen-

tu horietan agertuko diren "banku" hitzen adiera finantza-erakundeari dagokiona izango da ziurrenik, eta ez parkeko eserlekuari dagokiona. Honek desanbiguazioaren beharra gutxitzen du.

– Hitz askok adieren banaketa asimetrikoa dute; hots, hitz baten adiera bat hitz horren beste adierak baino askoz erabiliagoa da. Horrelako kasuetan, eta hitzek adiera bakarra dutenetan, desanbiguazioa alferrikakoa izango da, besterik ezean ere, adiera egokia dutelako hitz horien agerpenek. Hau frogatzeko erabili zuten datu-multzoko % 75,6 kontsultetan aipatutako bi kasu horietakoren bat gertatzen zen.

Sanderson-ek (1994) Krovetz eta Croft-ek (1992) esandakoak berretsi eta desanbiguazioko errore-tasak IBan zenbaterainoko eragina duen aztertu zuen. Horretarako, *sasihitz*en bidez jatorrizko dokumentuei anbiguotasuna gehitu zien. Lan honetatik ateratako ondorioak hauek ziren: desanbiguatzailea erabilgarria zela bakarrik galdera motzetan edo desanbiguatzailearen zehaztasun-maila oso altua zenean bakarrik (% 20-30 inguruko edo handiagoko errore-tasa baldin badu, hobe desanbiguatu gabe uztea).

Gonzalo *et al.*-ek (1998) eskuz desanbiguatutako dokumentuak erabili zituzten zuzen desanbiguatutako dokumentu-bilduma bat erabiliz berreskurapenean lor zitezkeen hobekuntzak zenbatekoak ziren ikusteko. Hainbat esperimentu egin zituzten, bakoitzean indize desberdin bat erabiliz: hitzetan oinarritutakoa, adieretan oinarritutakoa, eta, Wallis-ek (1993), Richardson eta Smeaton-ek (1995) eta Smeaton eta Quigley-k (1996) egin zuten antzera, baita WordNeteko *synset*etan oinarritutakoa ere. Adierak erabiliz hitzak soilik erabilita baino emaitza hobeak lortu zituzten galdera batzuetarako; ez, ordea, beste batzuetarako, eta hori erabili ziren adierak berreskurapenatazarako zehatzegiak zirelako gerta zitekeela ondorioztatu zuten. Emaitzarik onenak *synset*ak erabiliz lortu zituzten. Lan honetan ere, Sanderson-ek bezala, desanbiguazio-erroreek IB sistemaren eraginkortasunean zuten eragina neurtu zuten, eta % 40-50 bitartekoa baino handiagoko errore-tasadun desanbiguatzailearekin eraginkortasunak behera egiten zuela ikusi zuten. Sanderson-ek esandakoa baino errore-tasa altuagoa izanik, esan daiteke desberdintasun hori esperimentuen artean zeuden desberdintasunengatik izan daitekeela; esaterako, Sanderson-ek hitzen adierak erabili zituen eta beste lan honetan *synset*ak erabili zituzten.

Orain arte aipatu ditugun lanetan errepresentazio desberdinak erabili bazituzten ere, denetan hitzak adiera bakarraz ordezten ziren. Sanderson-ek (1997) hitzak bere adiera guztiekin ordeztu zituen, bakoitzari pisu bat esleituz. Baina, oraingoan ere, ez zituen emaitza onak lortu, eta desanbiguazio-

erroreengatik zela adierazi zuen. Emaitzetan bazegoen salbuespen bat, ordea: hitz bakarreko kontsultetan emaitza hobeak lortu zituen hitzen adierak erabiliz. Gainera, adiera bakarra beharrean, hitz baten adiera guztiak erabiliz emaitza hobeak lortzen zirela ikusi zuen.

Jarraian aipatuko ditugun bi lanetan, aurrekoetan ez bezala, berreskurapenerako erabiliko zen dokumentu-bilduma bera erabili zuten HAD sistema sortzeko.

Zernik-ek (1991) zioen hiztegiko adieren definizioek ez zutela balio IB atazetarako, adiera-banaketa finegia zelako eta semantikan baino, gramatika-irizpideetan gehiago oinarritzen zirelako. Horrela bada, berreskurapenerako erabiliko zituen dokumentuetako hitzen agerpenak testuinguruko hitzen arabera multzokatu, eta, ondoren, multzo bakoitza hiztegiko adieraren batekin lotzen saiatu zen. 30 hitz desanbiguatu zituen eta ez zuen bere HAD sistemaren zehaztasun-neurririk eman. Hitz horiek desanbiguatuta egindako probetan galdera luzeetan ez zuen IBaren eraginkortasunean alderik nabaritu, hitz bakarrekoetan, ordea, bai.

Azken honen antzera, Schutze eta Pedersen-ek (1995) corpusean bakarrik oinarritutako HAD sistema bat eraiki zuten. Berreskurapeneko esperimentu txiki bat egin zuten —25 kontsulta— eta desanbiguatuta berreskurapen-sistemaren eraginkortasunean % 14ko hobekuntzak lortu zituzten. Lan hau da HAD sistema IB sistema batean txertatu eta emaitza positiboak aurkezten lehena.

Geroago ere emaitza positiboak lortu dituzten lan gehiago argitaratu dira. Jarraian aipatuko ditugun lan hauek doitasun handiz desanbiguatu daitezkeen hitzetan jartzen dute arreta batez ere.

Esate baterako, Mihalcea eta Moldovan-ek (2000) hitz guztiak desanbiguatu beharrean, zehaztasun handiz desanbiguatu ahal zituzten hitzak bakarrik desanbiguatu zituzten (izen eta aditzen % 55a), % 92 baino zehaztasun altuagoa lortuz. Desanbiguatutako corpus horretatik *synset* eta hitzetan oinarritutako indizeak erabiliz, hobekuntzak lortu zituzten IB esperimentuetan.

Krovetz eta Croft-ek (1992) esan zutena —hitzen kolokazioaren eragina eta adieren banaketa asimetrikoa— kontuan hartuz, Stokoe *et al.*-ek (2003) zehaztasun gutxiko desanbiguazioak ekidingo zituen desanbiguatzaile bat garatu zuten, honela: hitz baten adiera zehazteko nahikoa informazio ez baldin badago kolokazio (*collocation*) edo agerkidetzak (*co-occurrence*) kontuan hartuz, WordNeteko maiztasun handiena duen adiera esleituko zaio. Desanbiguatzaile honetan oinarrituta, adieratan oinarritutako berreskurapena egin,

eta hobekuntzak lortu zituzten. Hobekuntza horiek, ordea, artearen egoerako emaitzetara iristen ez zen oinarri-lerro sinple batekiko lortu zituzten.

Kim *et al.*-ek (2004) ere ildo beretik jarraitu zuten, eta desanbiguazio-erroreen eragina txikiagotzeko izenak bakarrik desanbiguatu zituzten, ale larriko (*coarse-grained*) desanbiguazioa deritzona aplikatuz —WordNeteko 25 erro-adiera (*root sense*) bakarrik erabiliz— eta hitz batzuei adiera bat baino gehiago esleituz. Horrela, artearen egoerako IB sistemen eraginkortasuna hobetzea lortu zuten.

Liu *et al.*-ek (2004) WordNeten oinarritutako doitasun handiko desanbiguazio-algoritmo baten bidez kontsultako terminoak desanbiguatu eta, kasu batzuetan, hedapena eginez berreskurapeneko emaitzak hobetzea lortu zuten —kontsulta motzekin probatu zuten. Lan honen jarraipenean (Liu *et al.*, 2005) HAD sistema gehiago garatu zuten, WordNetez gain, weba ere erabiliz desanbiguatzerako garaian. Kontsulta motzak desanbiguatzen % 90eko zehaztasuna lortu zuten, eta, desanbiguatutako kontsulta horiek erabiliz, hainbat datu-multzorekin ordura arteko emaitzarik onenak lortu zituzten.

Handik urte batzuetara, ikerketa-lerro hau gehiago jorratu nahian, CLEF[6] ebaluazio-kanpainaren baitan Robust-WSD ataza antolatu zen (Agirre *et al.*, 2009a, 2010d). Ataza honen helburua HADak berreskurapenean zenbaterainoko ekarpenak egin zitzakeen ikustea zen. Horretarako, atazako parte-hartzaileei automatikoki desanbiguatutako datu-multzo bat —WordNeteko *synset*ez zegoen etiketatua— ematen zitzaien beraien IB esperimentuak egin zitzaten. Emaitzetan ez zen joera garbi bat ikusi, emaitzarik onenak desanbiguazioko informazioa erabili gabe lortu baziren ere, parte-hartzaileetako batzuek lortu baitzuten beraien emaitzak hobetzea informazio hori erabiliz. Adibidez, Pérez-Agüera eta Zaragoza-k (2009) kontsulten hedapena egiterakoan, terminoen arteko antzekotasun semantikoa kontuan hartzea proposatu zuten, horren araberako kontsulta egituratuak sortzeko. Antzekotasun semantikoa kalkulatzeko WordNet erabili zuten, eta kontsultako terminoak WordNeteko kontzeptuekin lotzeko HADeko informazioa erabili zuten. Kontsulta egituratu horiek erabiliz, artearen egoerako oinarri-lerroko sistemaren emaitzak hobetzea lortu zuten. Guk ere bi alditan parte hartu genuen ataza honetan (Agirre *et al.*, 2009b, 2010e); lehenengoan HAD informazioa erabili genuen hedapenak egiteko, eta bigarren aldian HAD informazioaz gain, ahaidetasun semantikoaz baliatu ginen hedapenak egiteko. Esperimentu hauen inguruko berri gehiago esperimentuei dagozkien kapituluetan emango dugu.

[6] http://www.clef-campaign.org/

Lan berrienekin jarraituz, Giunchiglia *et al.*-ek (2009) egin zuten hastapeneko lanean kontsulta eta dokumentuetako hitzak ahal zenean WordNeteko adierekin ordezkatu —horretarako HADa erabiliz—, eta adieretan oinarritutako errepresentazio horiek erabiliz, berreskurapeneko emaitzak hobetzea lortu zuten.

Ikusi dugun bezala, HADa berreskurapen-sistema batean modu egokian txertatzeko ahaleginak ugariak izan dira. Laburbilduz, hauetako gehienetan desanbiguazioaren bitartez hitz bakoitzaren adiera egokia (edo bat baino gehiago kasu batzuetan) aukeratu eta adiera horietan oinarrituta bilaketak egiten dituzte, eta ez dituzte hitzak edo hitzen erroak erabiltzen bilaketetan. Liu *et al.*-ek (2004), behin kontsultako hitz bakoitzaren adiera egokia zein zen jakinda, hedapena egin zuten, eta hedapenetik lortutako hitzak (edo hitz-erroak) erabiliz egiten zituzten bilaketak. Aipatu ditugun lanetan HADa egiteko kanpoko ezagutza-iturriren bat erabiltzen dutenen artean, WordNet ontologia da erabiliena. Guk ere, 4. kapituluan azalduko ditugun esperimentuetan, Liu *et al.*-ek (2004) egin zutenaren antzera, HADeko informazioa erabili dugu kontsulten hedapena egiteko; baina, horretaz gain, dokumentuen hedapena ere egiten dugu. Horretarako, aurrez WordNet erabiliz desanbiguatuta zegoen datu-multzo bat erabili dugularik.

Antzekotasun semantikoa erabiliz

Parekatze-arazoari aurre egiteko HADa erabiltzeaz gain, antzekotasuna ere erabil daiteke. Antzekotasuna, ordea, oso kontzeptu zabala da, bere baitan antzekotasun mota desberdinak biltzen dituena. IBaren arloan txertatzerakoan ere, hainbat hurbilpen daude antzekotasunera iristeko; besteak beste, estatistiketan oinarritutako antzekotasuna eta antzekotasun semantikoa[7]. Gainera, unitate desberdinei aplikatu ahal zaie; alegia, hitzen arteko edo dokumentuen arteko antzekotasuna erabil daiteke. Horrela, adibidez, erlazionatutako hitzak eta dokumentuak oso erabiliak izan dira hainbat lanetan, jarraian ikusiko dugun moduan.

Voorhees-ek (1994) antzekotasun semantikoan oinarrituz, kontsulta bakoitzari harekin erlazionatutako WordNeteko *synset*ak gehitu zizkien eskuz, eta ondoren, *synset* horietatik abiatuz hedapena egin zuen. Kontsultako jatorrizko termino eta hedapenetik lortutako terminoak erabiliz hobekuntzak

[7] Askotan antzekotasun semantikoa eta ahaidetasun semantikoa nahastu egin ohi dira, eta lehena erabili ohi da bigarrena erreferentziatzeko. Bi termino hauen arteko desberdintasunaz 3.4.2 atalean hitz egingo dugu.

lortu zituen ebaluazioan erabilitako kontsulta motzenetan.

Kontsulten hedapenean ez ezik, dokumentuen hedapenean ere erabil daiteke antzekotasuna. Dokumentuen hedapena lehen aldiz hizketaren berreskurapenean (*speech retrieval*) erabili zen. Hizketa-transkripzioak zarata handiko dokumentuak izanik, Singhal eta Pereira-k (1999) hauetako dokumentu bakoitzaren errepresentazioarekin erlazionatutako dokumentuak gehitzea proposatu zuten. Erlazionatutako dokumentu horiek lortzeko, beste corpus batean berreskurapena egiten zuten jatorrizko dokumentua kontsulta bezala erabiliz, eta berreskurapenean lortutako lehen hamar dokumentuak hartzen zituzten.

Beste alde batetik, dokumentu-hedapen modura ikus daiteke dokumentuak multzokatzea (*document clustering*) ere. Izan ere, antzeko dokumentuak multzokatuz berreskurapena egiten denean, dokumentu bat berreskuratu ahal izango da, nahiz eta kontsultako terminoak ez agertu dokumentu horretan —baldin eta multzoko beste dokumentuetan baldin badaude termino horiek. Dokumentuen multzokatzea berreskurapenean erabiltzeko ahaleginak aspaldikoak dira (Jardine eta van Rijsbergen, 1971; Salton, 1971a; Croft, 1978; Voorhees, 1985; Can eta Ozkarahan, 1990). Lan horietan eta beste hauetan, (Can *et al.*, 2004), (Singitham *et al.*, 2004) eta (Altingövde *et al.*, 2008), dokumentuen multzokatzearen bidezko berreskurapenaren eraginkortasun eta errendimenduak aztertu zituzten, eta emaitza nahasiak izan ziren: batzuek dokumentuak multzokatuz hobekuntzak lortu zituzten; beste batzuek, aldiz, ez. Beste lan batzuk aipatzearren, Kurland eta Lee-k (2004) eta Liu eta Croft-ek (2004) hizkuntza-ereduen gainean dokumentuen multzokatzea aplikatu eta berreskurapeneko emaitzetan hobekuntzak lortu zituzten.

Azken lan horretan dokumentuen multzokatzea, besteak beste, hizkuntza-ereduen leuntzea (*smoothing*) hobeto egiteko erabiltzen dute. Azken urteetan indar handia hartu dute hizkuntza-ereduek IB sistemetan, eta ikerketa-lanek diotenez, garrantzi handia du leuntzea ondo egiteak. Hori dela eta, azken urteetan hainbat lanetan leuntzea hobeto egiteko dokumentu edo hitzen arteko antzekotasunak modu ezberdinetara erabili dituzte: dokumentuen multzokatzearen bidez (Liu eta Croft, 2004), dokumentuen hedapenaren bidez (Tao *et al.*, 2006) eta hitz-grafoen bidez (Mei *et al.*, 2008; Huang *et al.*, 2009). Denek ere beraien ekarpenekin emaitza positiboak lortu zituzten.

Gure esperimentu batzuetan ere, grafoak erabiliko ditugu. Baina, ez dute oraintsu aipatutako lan hauen antz handirik. Izan ere, aipatu berri ditugun lan hauek antzekotasuna erabiltzen badute ere, agerdiketzan edo bestelako

neurri estatistikoren batean oinarritutako antzekotasuna erabiltzen dute; eta ez antzekotasun semantikoa, gure lanean bezala.

Aurreko atalean aipatutako lan batzuetan, HADa erabiltzeaz gain, antzekotasun semantikoa ere erabiltzen zuten. Zehaztuz, Richardson eta Smeaton-ek (1995) kontsulta eta dokumentuen antzekotasun semantikoan oinarritzen zen berreskurapen-eredua erabili zuten; Pérez-Agüera eta Zaragoza-k (2009) jatorrizko kontsultako terminoen eta hedapenean sortutako terminoen arteko antzekotasun semantikoaren arabera sortzen zituzten kontsulta egituratuak kontsultaren hedapen-prozesuan.

Beste lan bat aipatzearren, Varelas *et al.*-ek (2005) kontsultaren hedapeneko terminoak antzekotasun semantikoaren arabera aukeratzen zituzten, eta Richardson eta Smeaton-ek (1995) egin zutenaren antzera, kontsulta (kasu honetan kontsulta hedatua) eta dokumentuen arteko antzekotasunean oinarrituz egiten zen berreskurapena. Webeko irudi eta dokumentuen berreskurapenean emaitza itxaropentsuak lortu zituzten. Esan daiteke gure esperimentuetan ere, azken lan honetan bezala, kontsultaren (eta dokumentuaren) hedapeneko terminoak antzekotasun semantikoaren arabera aukeratzen direla. Baina gure esperimentuetan, beste horiek ez bezala, antzekotasuna kalkulatzeko WordNeteko hitz eta erlazio guztiak erabili ditugu, eta grafo-algoritmo bat erabili dugu.

Ikusi ditugun lan askotan antzekotasuna lortzeko WordNet ezagutza-basea erabili da. Ezagutza-basean oinarritutako IBa aspaldidaniko kontua da. Historian zehar, IBaren beharra liburutegietan antzeman zen, eta han egin ziren lehen berreskurapenak. Urte haietan liburutegietan egiten ziren bilaketak thesaurus espezifiko bat erabiliz egiten ziren. Horrela, bilaketa hiztegi edo hitz-zerrenda batera mugatzen zen —indizean ere hitz-zerrenda horretako hitzak bakarrik zeuden—, eta bilaketa on bat egitea errazagoa zen. Gerora bestelako ontologiak edo ezagutza-base konplexuagoak erabiltzen hasi ziren, ez bakarrik indize eta bilaketak hiztegi konkretu batera mugatzeko, baita kontsulta eta dokumentuak aberasteko, adibidez, hedapenak eginez. Edozein modutan ere, domeinu konkretu batera mugatutako ontologia edo ezagutza-baseak erabiltzen jarraitu da —energia-teknologia arloan (Monarch eta Carbonell, 1987), medikuntza arloan (Rada *et al.*, 1989), etab.—. Orain arte nagusitu den domeinu orokorreko ezagutza-basea WordNet da, baina azkenaldian joera aldatuz doa, eta gero eta lan gehiago dira, adibidez, Wikipedia ustiatzen dutenak (Gabrilovich eta Markovitch, 2009).

Ezagutza-base hauek duten egitura dela eta —kontzeptuen hierarkia-erlazioez lotuak daudenak—, grafo moduan errepresentatzeko aukera ematen

dute, eta, ondorioz, grafoei aplikatzen zaizkien algoritmoekin tratatu. Adibidez, *aktibazioaren hedatze* delakoa (ingelesez *spreading activation* moduan ezaguna dena) IBaren arloan erabiltzea aspaldian proposatu zuten (Salton eta Buckley, 1988). Lan berrienetan ere halako teknikak erabiltzen dira. Esaterako, Hsu *et al.*-ek (2008) kontsultaren hedapenerako terminoak lortzeko WordNet eta ConceptNet ezagutza-baseei aktibazioaren hedatzea aplikatzen zien. Lan honen antzera, guk ere aktibazioaren hedatzearekin erlazionatuta dagoen *ausazko bide* izeneko algoritmoa (*random walks*) (Hughes eta Ramage, 2007) erabiliko dugu WordNeten grafotik kontsulta eta dokumentuak hedatzeko terminoak lortzeko. Erabilitako algoritmo hori bera hitzen antzekotasun semantikorako (Agirre *et al.*, 2009c) eta HADa egiteko (Agirre eta Soroa, 2009) erabilia izan da emaitza arrakastatsuekin.

Egungo egoera

Ikusi dugun moduan, saiakera asko egin dira HADa eta ahaidetasun semantikoa erabiliz parekatze-arazoari aurre egiteko, beti ere IBko emaitzak hobetu nahian. Hasierako lanetan orokorrean emaitza negatiboak lortu bazituzten ere, azkenaldian emaitza onak eta itxaropentsuak lortu dira.

Aipatu dugu lan batzuetan IB sistemen eraginkortasuna hobetzeko HAD edo beste teknikaren bat erabilgarria izatea (edo ez izatea) erabiltzen den HAD sistemaren edo beste teknika horren zehaztasunaren menpe egon daitekeela. Ikus dezagun bada zein den lexiko-semantikako sistema hauen artearen egoera. Alde batetik, gaur egungo HAD sistema hoberenek % 82-83ko zuzentasuna lortzen dute ale larriko (*coarse grained*) desanbiguazio-atazetan (Navigli, 2009). Bestetik, dokumentuen arteko ahaidetasuna kalkulatzen duen sistema batek orain arte lortu duen Pearson-en korrelazio-koefizienterik altuena 0,72koa da (Gabrilovich eta Markovitch, 2009).

Lan honetan, aipatutako bi teknika hauek erabiliz, IB sistemaren eraginkortasuna hobetzen ahalegindu gara.

3 Esperimentazio-ingurunea

Tesi-txosten honetan azalduko ditugun esperimentuak garatzeko esperimentazio-ingurunea deskribatu eta hurrengo kapituluetan kontatuko duguna ulertzeko hainbat oinarrizko kontzeptu azalduko ditugu kapitulu honetan. Alde batetik, IBko oinarrizko ataza nola egikaritzen den eta horretarako erabil daitezkeen algoritmoak eta tresnak zeintzuk diren azalduko dugu. Bestetik, gure esperimentuetan erabilitako hizkuntza-prozesamenduarekin zerikusia duten hainbat teknika eta baliabide deskribatuko ditugu. Horretaz gain, esperimentuak egiteko erabilitako datu-multzoak aurkeztu eta ebaluazioaren inguruko hainbat xehetasun emango ditugu.

3.1 Informazioaren berreskurapenerako ad hoc ataza

Informazio-bilaketa baten atzean hainbat arrazoi eta helburu egon daitezke, eta horren arabera, berau gauzatzeko beste hainbeste bide edo modu. Adibide ohikoenak hauek izan daitezke (Meij, 2010): erabiltzaileak buruan duen web-orri edo dokumentu jakin bat bila dezake (ataza honi ingelesez *named-page finding* deritzo), edo aditu edo entitate adierazgarriak bilatu nahiko ditu (*expert/entity finding* ataza), edo galdera zehatz baten erantzuna jakin nahiko du (*galderak erantzutea* edo *question answering* ataza), edo gai baten inguruko informazioa aurkitu nahiko du. Azken hau ingelesez *topic-finding* edo *ad hoc retrieval* moduan ezagutzen da, eta IBko atazarik arruntenetarikoa da. Tesi-lan honetan aurkeztuko ditugun IBko esperimentu guztiak mota

honetakoak dira.

Ad hoc ataza batean erabiltzaileak informazio-behar bat du (*information need*) eta behar hori asetzeko bilaketa bat egiten du dokumentu-bilduma baten gainean IB sistema bat erabiliz. Sistema honek kontrolpean duen dokumentu-bilduma horretatik informazio-behar horrentzako adierazgarriak diren dokumentuak itzuliko dizkio erabiltzaileari. Ikus ditzagun, banan-banan, horrelako ataza bat egikaritu ahal izateko beharrezko elementuak:

- **dokumentu-bilduma**: IB sistemak bilduma osatzen duten dokumentuetan egingo ditu bilaketak.

- **indizea**: Dokumentu-bilduman agertzen den hitz bakoitza zein dokumentutan agertzen den gordetzen duen fitxategia (dokumentuko posizioak edo agerpen kopuruak ere gordetzen dira batzuetan). Bilaketak modu azkar eta eraginkor batean egitea ahalbideratzen du indizeak.

- **informazio-beharra**: Erabiltzaileak gai baten inguruan gehiago jakin nahi duelako egiten du bilaketa, eta bilaketa hori egiteko erabiltzaileak adierazten duen hori da informazio-beharra.

- **kontsulta**: Kontsulta bat erabiltzaileak bere informazio-beharra sistemari adierazteko erabiliko duen hitz-segida da.

- **bilaketa-teknika(k)**: IB sistemak kontsulta dokumentuekin parekatzeko bilaketa-teknika bat (edo gehiago) erabiliko du.

- **berreskuratutako dokumentuak**: Bilaketa-teknikak egiten duen parekatzean IB sistemak dokumentu bakoitzari pisu bat esleituko dio dokumentuaren adierazgarritasuna adieraziz. Informazio-beharra asetzeko dokumentu horretako informazioa zenbateraino den baliagarria esaten du adierazgarritasun-pisu horrek. Pisu horren arabera ordenatu (adierazgarriena lehen posizioan jarriz) eta lehen k dokumentuen rankinga izango da berreskuratutako dokumentuen zerrenda.

- **adierazgarritasun-epaiak**:[1] Adierazgarritasun-epaietan informazio-behar bakoitzerako zein dokumentu diren adierazgarriak (eta zeintzuk ez) zehazten da. Dokumentu bat adierazgarria dela esango dugu baldin eta dokumentu horretako informazioa baliogarria bada dagokion

[1] Izen hau ingeleseko *relevance judgments*etik itzuli dugu.

informazio-beharra asetzeko. Gure esperimentuetan erabili ditugun adierazgarritasun-epaietan adierazgarritasuna beti modu bitarrean adierazten da: adierazgarria da, edo ez da adierazgarria. Berreskuratutako dokumentuak ebaluatzeko epai hauek erabiliko dira, berreskuratutako dokumentuen zerrenda adierazgarritasun-epaien zerrendarekin konparatuz.

Labur azalduta, honela egikarituko litzateke ad hoc ataza arrunt bat. Lehenbizi, eta behin bakarrik egitea nahikoa da, IBko tresna bat erabiliz, dokumentu-bildumarekin indizea sortzen da. Ondoren, informazio-behar bakoitzerako ondorengo ziklo hau errepikatuko da behin eta berriz: (i) informazio-behar horretan oinarrituta, kontsulta bat sortzen du giza erabiltzaileak[2]; (ii) IB sistemako algoritmo bat (edo gehiago) erabiliz, eta kontsulta horretan oinarrituz, dokumentu batzuk berreskuratuko dira; (iii) berreskuratutako dokumentu horiek ebaluatuko dira adierazgarritasun-epaiek diotenaren arabera.

IB algoritmo edo teknika bat ebaluatzea denean helburua (gure kasuan bezala), kontsulta asko egin behar dira bata bestearen atzetik. Horrelako kasuetan ataza nola egikaritzen den (ebaluazioko exekuzio edo *run* bat zer den) 3.7 atalean azalduko dugu.

3.2 Informazioaren berreskurapenerako algoritmoak

IB sistema baten helburua kontsulta baten araberako dokumentuak berreskuratu eta hauek modu egokian ordenatzea da: dokumentu adierazgarriek ez-adierazgarrien aurretik egon behar dute. Hau lortzeko, sistemak dokumentuei pisu bat esleitu behar die, eta dokumentu adierazgarriek pisu handiagoa izan beharko lukete ez-adierazgarriek baino (Zhai, 2008).

Ranking-funtzioen ardura da dokumentuei pisuak esleitzea. Berreskurapen-funtzioa adierazgarritasuna formalizatzen duen berreskurapen-eredu batean oinarritzen da. Berreskurapen-eredu asko garatu eta probatu dira

[2]Ohikoena informazio-beharra lengoaia naturalean adierazitako esaldi arrunt bat (edo gehiago) izatea da. Kontsulta esaldi horretako gako-hitzak diren terminoak hartuz osatuko da. Batetik besterako prozesu honetan *stopword*-zerrendak eta lematizatzaile edo *stemmer*ak erabil daitezke (argibide gehiago 3.4.5 atalean)

urteetan zehar, baina ez da aurkitu bat beste guztiak baino eraginkorragoa denik.

Gure esperimentuetan oinarri-lerro moduan emaitza oso onak eman dituzten bi eredu desberdin erabili ditugu: BM25 eredu probabilistiko klasikoa, eta hizkuntza-eredu probabilistikoan oinarritutako *query likelihood* eredua, euskaraz *kontsulta-egiantza* eredua deituko dioguna. Eredu hauetan oinarrituz inplementatu ditugu semantika kontuan hartzen duten gure algoritmoak ere.

Jarraian, bi oinarri-lerro hauek eta IBan oso erabilia den *sasiadierazgarritasun-feedback* edo *pseudo-relevance feedback* deitzen den teknika azalduko ditugu.

3.2.1 BM25

BM25 ranking-funtzio bat da, bilatzaileek testu-dokumentuak berreskuratu eta ordenatzeko erabiltzen dutena. *Okapi BM25* izenarekin ere ezagutzen da, Okapi testuaren berreskurapenerako sistemarako inplementatu baitzen lehenengo aldiz.

Eredu probabilistikoan oinarritzen da, eta, hitz gutxitan esanda, dokumentuak kontsultarekiko duten adierazgarritasunaren arabera ordenatzen ditu (Robertson eta Walker, 1994). Kontsultako terminoak dokumentuetan zenbat aldiz agertzen diren eta dokumentuen luzerak (hitzetan kontatuta) kontuan hartzen ditu. Honela definitzen dute Robertson eta Zaragoza-k (2009) termino baten BM25 pisua:

$$w_{Dt}^{\text{BM25}} = \frac{\text{tf}_{tD}}{k_1 \left((1-b) + b\frac{l_D}{l_{bb}} \right) + \text{tf}_{tD}} \text{idf}_t \tag{3.1}$$

non tf_{tD} D dokumentuan t terminoaren agerpen kopurua den, l_D D dokumentuaren luzera den (hitz kopurutan neurtua), l_{bb} dokumentu-bildumako dokumentuen batez besteko luzera den, idf_t dokumentu-maiztasunaren alderantzikatua den [3] (edo zehatzago esanda RSJ pisua) (Robertson eta Zaragoza, 2009, 3.4.5 atala), eta k_1 eta b parametro askeak diren.

[3] *idf*, *inverse document frequency*ren laburdura da. Pisu hau IB arloan asko erabiltzen den neurri estatistiko bat da, dokumentu-bilduma batean asko agertzen den termino baten eragina gutxitzeko. Alegia, bilduman oso gutxi agertzen den termino baten *idf* pisua altua izango da, eta, alderantziz, oso ohikoa den termino baten *idf* pisua baxua izango da. Pisu hau kalkulatzeko hainbat aldaera baldin badaude ere, hemen, MG4J tresnak nola

Parametro aske hauekin funtzioa esku artean darabilgun datu-multzora egokitu daiteke. k_1 parametroarekin ($k_1 > 0$) funtzioaren amaierako pisuan dokumentuko terminoaren maiztasunak izan dezakeen eragina kontrola dezakegu. Horrela, esaterako, $k_1 = 0$ bada, terminoaren agerpen kopurua ez da kontuan hartzen. b parametroa doituz ($0 \leq b \leq 1$) dokumentu luzerak izango duen eragina erabaki dezakegu. $b = 0$ bada, dokumentu luzerak ez du amaierako pisuan eraginik izango, eta, aldiz, $b = 1$ bada, luzeraren normalizazio osoa egingo da ($\frac{l_d}{l_{bb}}$). Nahiz eta hainbat esperimenturen arabera $0,5 < b < 0,8$ eta $1,2 < k_1 < 2$ balio egokiak izan parametro hauentzat, beste hainbat faktorek ere eragin dezakete hauen balio optimoetan, eta horregatik oso komenigarria da kasu bakoitzean parametro hauek doitzea.

Esan bezala, goiko funtzio honen bidez, termino baten pisua kalkulatzen da. Hortaz, dokumentu osoaren pisua kontsultako termino guztien w_{Dt}^{BM25} pisuak batuz kalkulatzen da.

BM25 ranking-funtzioa asko erabili da IBko hainbat atazatan oso emaitza onak lortuz (Manning *et al.*, 2009). Gainera, hainbat IB sistema libretan ere inplementatuta dago. Guk 5. kapituluko esperimentuetan funtzio hau erabili dugu, MG4J sistemaren bitartez (ikus 3.3.2 atala). MG4J sistemak 3.1 ekuazioan zehaztutakoaren aldaera bat dauka inplementatuta.

3.2.2 *Query likelihood* eredua

Query likelihood (QL) edo kontsulta-egiantza delakoa IBrako hizkuntza-eredu probabilistiko bat da[4]. Eredu honetan bildumako D dokumentu bakoitzaren Θ_D hizkuntza-eredua sortzen da. Eta Θ_D estimazio horretan oinarrituta, D dokumentutik Q kontsulta sortzeko egiantzaren arabera ordenatzen dira dokumentuak berreskuratze-prozesuan. Hortik datorkio kontsulta-egiantza izena eredu honi.

kalkulatzen duen zehaztuko dugu:

$$\text{idf}_t = \log \frac{N - \text{df}_t + 0,5}{\text{df}_t + 0,5}$$

non N bildumako dokumentu kopurua den eta df_t t terminoa agertzen den dokumentu kopurua den.

[4]Hizkuntza-eredua, 2.1 atalean esan bezala, hitz-segidei probabilitate-banaketak esleitzen dizkien eredu probabilistiko bat da. Sinpleena unigrametako hizkuntza-eredua da, non estimazioak egiterakoan hitz bakoitza besteekiko independentetzat hartzen den. Hizkuntzaren prozesamenduan asko erabiltzen da, eta baita IBan ere.

Ikus dezagun jarraian zehatzago. Eman dezagun Q kontsulta bat izanik $P(D \mid Q)$ probabilitatearen arabera ordenatu nahi ditugula dokumentuak. Probabilitate horren estimazioa egiteko aukera bat Θ_Q kontsultaren hizkuntza-eredutik dokumentua sortzeko probabilitatea ($P(D \mid \Theta_Q)$) kalkulatzea da. Baina, kontsulta gehienak nahiko motzak izaten direnez, ez genuke testu gehiegi edukiko Θ_Q hizkuntza-ereduaren estimazioa egiteko. Horren ordez, beste bide bat jarraitu ohi da, hain zuzen ere, jarraian ikusiko duguna.

Bayes-en teorema jarraituz, honela geldituko zaigu kalkulatu nahi dugun probabilitatea:

$$P(D \mid Q) = \frac{P(Q \mid D)P(D)}{P(Q)} \tag{3.2}$$

$P(Q)$ berdina da dokumentu guztientzat, eta, hortaz, kontsulta bat emanik, ez du dokumentuen rankingean eraginik. $P(D)$ dokumentua aukeratzeko aldez aurretiko probabilitatea da. Probabilitate hau dokumentuaren luzera, berritasuna, aurretik zenbat pertsonak irakurri duten, edo beste hainbat faktoreren araberakoa izan badaiteke ere, guk horrelakorik kontuan izango ez dugunez, dokumentu guztientzat probabilitate hori uniformea dela kontsideratuko dugu. Hortaz, P(Q) eta P(D) alde batera utziko ditugu, eta hau izango dugu:

$$P(D \mid Q) \overset{\text{ordena}}{=} P(Q \mid D) \tag{3.3}$$

Hortaz, dokumentuak $P(Q \mid D)$ probabilitatearen arabera ordenatuko dira. Alegia, lehen esan bezala, D dokumentuaren hizkuntza-eredutik Q kontsulta sortua izateko probabilitatearen arabera ordenatuko dira dokumentuak. Hara nola azaltzen duten Manning *et al.*-ek (2009) zein den eredu honen atzean dagoen ideia:

> "The intuition of the basic model is that the user has a prototype document in mind, and generates a query based on words that appear in this document. Often, users have a reasonable idea of terms that are likely to occur in documents of interest and they will choose query terms that distinguish these documents from others in the collection."

D dokumentuaren hizkuntza-eredutik (Θ_D) Q kontsulta sortua izateko probabilitatearen estimazioa egiteko hizkuntza-eredu multinomiala aukeratu dugu eta hitz guztiak independentetzat hartu (Metzler eta Croft, 2004):

$$P(Q \mid \Theta_D) = \prod_{i=1}^{|Q|} P(q_i \mid \Theta_D)^{\frac{1}{|Q|}} \tag{3.4}$$

non Q kontsultako terminoa den q_i, eta $|Q|$ kontsulta horren luzera den (termino kopurua). $P(q_i \mid \Theta_D)$ definitzeko Dirichlet leuntze-teknika (Zhai eta Lafferty, 2001a) jarraituz, honako hau dugu:

$$P(q_i \mid \Theta_D) = \frac{\mathrm{tf}_{q_i D} + \mu \frac{\mathrm{tf}_{q_i C}}{|C|}}{|D| + \mu} \tag{3.5}$$

non $\mathrm{tf}_{q_i D}$ eta $\mathrm{tf}_{q_i C}$ D dokumentuko eta bilduma osoko q_i kontsulta-terminoaren maiztasunak diren, hurrenez hurren, eta μ leuntze-teknikaren parametro askea den.

Esan beharra dago hizkuntza-ereduetan leuntze-tekniken erabilera oso garrantzitsua dela. Izan ere, ez bagenu horrelakorik erabiliko, $\Theta_D$tik kontsulta-termino bat sortzeko probabilitatea egiantza handieneko estimatzailearen (*maximum likelihood estimation*) arabera, honako hau izango litzateke:

$$P(q_i \mid \Theta_D) = \frac{\mathrm{tf}_{q_i D}}{|D|}$$

Baina, kasu honetan, dokumentuan agertzen ez den kontsulta-termino baten probabilitatea 0 izango zen. Eta, hortaz, kontsulta osoaren $P(Q \mid \Theta_D)$ ere 0 izango zen, terminoetako bat dokumentuan ez egoteagatik. Horrelakoak ekiditeko, hots, zero-probabilitateak ekiditeko eta dokumentuan agertzen ez diren terminoei probabilitate-masa txiki bat esleitzeko, leuntze-teknikaren bat erabili ohi da. Gainera, dokumentuan behin azaltzen den hitzak probabilitate-masa handiegia izan ohi du, kasu askotan zoriz agertzen delako hitz hori dokumentu horretan. Horregatik, leuntze-teknikek dokumentuko agerpen kopuruak ez ezik, dokumentu-bilduma osoko maiztasunak ere hartzen dituzte kontuan (Manning *et al.*, 2009).

Zhai-ren (2008) arabera, QL berreskuratze-funtzioak Dirichlet leuntze-teknikarekin eta BM25 algoritmoak duten errendimendua parekoa da. Gure esperimentuetan hizkuntza-ereduak ere probatu nahi izan ditugu, eta 4. eta 6. kapituluetako esperimentuetan QL erabili dugu, hain zuzen ere, Indri sistemaren bitartez (ikus 3.3.1 atala).

3.2.3 *Pseudo-relevance feedback* metodoa

Informazioaren berreskurapena egitean nahiko ohikoa izaten da bilatu nahi ditugun dokumentuak ez lortzea lehenengo saiakeran, dokumentu horretako eta guk egindako kontsultako hitzak ez datozelako bat. Arazo horri aurre egiteko egindako kontsulta hori findu dezake erabiltzaileak, kontsultari hitz berriak gehituz.

Baina, IB sistemak ere lagundu dezake finketa horretan, erabat automatikoki sortuz kontsulta berria edota erabiltzaileari utziz zenbait aukeraketa egiten. Esaterako, *adierazgarritasun-feedback*a izeneko metodoan (ingelesez *relevance feedback* moduan ezagutzen dena) sistemak erabiltzailearen laguntzarekin hasierako kontsulta fintzen du amaierako emaitzak hobeak izan daitezen (Manning *et al.*, 2009). Zehatzago, hau da metodo honen prozesua:

(i) erabiltzaileak kontsulta sinplea egiten du;
(ii) sistemak dokumentuak berreskuratzen ditu;
(iii) erabiltzaileak dokumentu horietako batzuk adierazgarri edo ez-adierazgarri moduan markatzen ditu;
(iv) sistemak, erabiltzailearen feedback horretan oinarrituta, hasierako kontsulta hori fintzen du; eta
(v) kontsulta berri hori erabiliz, sistemak dokumentuak berreskuratzen ditu.

Metodo honen helburua, azken pauso horren ondoren lortutako dokumentuak (ii) pausoan lortutakoak baino hobeak izatea da. Izan ere, erabiltzailearentzat zaila da dokumentu-bilduma ikusi gabe hasierako kontsulta egoki bat egitea. Baina, erraz esan dezake sistemak itzulitako dokumentuetako batzuk adierazgarriak diren edo ez. Eta informazio hori sistemarentzat oso erabilgarria izan daiteke hasierako kontsulta hori baino hobea den beste kontsulta bat sortzeko.

Prozesu horretan erabiltzaileak parte hartzen duela ikusi dugu, dokumentu batzuen adierazgarritasuna zehaztuz. Baina, badago metodo honen aldaera bat non sistemak automatikoki egiten duen prozesu osoa, erabiltzailearen esku-hartzerik gabe. Metodo automatiko hau sasiadierazgarritasun-feedbacka da, ingelesezko *pseudo-relevance feedback* (PRF) izenez ezaguna dena (hori dela eta, hemendik aurrera txosten honetan zehar PRF moduan erreferentziatuko dugu). PRFaren prozesua oraintsu ikusitakoaren antzekoa da. Desberdintasun bakarra (iii) pausoan dago: erabiltzaileak esan beharrean zein dokumentu diren adierazgarriak eta zeintzuk ez, sistemak (ii) pausoan itzulitako dokumentu-zerrenda horretako lehen k dokumentuak adierazgarri-

tzat hartuko ditu. Eta k dokumentu horietan oinarrituko da termino berriak lortzeko.

PRFak, kontsultaren finketa edo hedapena egiteko modu automatiko bat den heinean, zarata sar dezake kontsultan, baina, artearen egoera aztertu dugunean aipatu dugu oso emaitza onak lortu direla berau erabiliz IB lan batzuetan. Aipatu, batez ere, sistemaren estaldura hobetzen dela metodo honekin. Horrexegatik, gure esperimentu batzuetan oinarri-lerro izateko aukeratu dugu metodo hau ere.

3.3 Informazioaren berreskurapenerako tresnak

3.1 atalean ikusi ditugu zein diren ohiko IB ataza baten pausoak. Indizeak sortu eta berreskurapena egiteko hainbat sistema daude merkatuan. Sistema hauek aurreko atalean ikusi ditugunak bezalako oinarrizko algoritmoak izan ohi dituzte inplementatuta. Sistemetako batzuk libreak direnez, bakoitzak bere ekarpenak ere inplementa ditzake oinarrizko algoritmo horien gainean.

Guk gure esperimentuetarako ondoren deskribatuko ditugun bi sistema libre erabili ditugu: Indri eta MG4J.

3.3.1 Indri

Indri testu-bilatzaile bat da, dokumentu-bilduma oso handiekin lan egin dezakeena. Hainbat dokumentu-formatu onartzen ditu, besteak beste, PDF, HTML edota XML formatuak.

Kontsulta egituratuak sortzeko kontsulta-lengoaia aberatsa du. Honela, eremu- edo pasarte-berreskurapena egiteko bidea ematen du. Gainera, kontsulta-terminoei pisuak emateko aukera eskaintzen du.

Hizkuntza-eredu (Ponte eta Croft, 1998) eta inferentzia-sareen (Turtle eta Croft, 1991) konbinazioan oinarritua dago bere baitako berreskurapen-eredua. Eredu konbinatu hau Metzler eta Croft-ek (2004) aurkeztu zuten, baina, ordutik hona hedatzen eta garatzen jarraitu dute eta tresnari buruzko informazio gaurkotua webgune honetan aurki daiteke: `http://www.lemurproject.org/indri.php`. Bilatzaile hau Lemur proiektuaren[5] baitan garatzen ari dira eta kode irekikoa da.

[5] http://www.lemurproject.org/

4. eta 6. kapituluetako esperimentuak egiteko tresna hau erabili dugu.

3.3.2 MG4J

MG4J (*Managing Gigabytes for Java*) testu-bilatzaile librea da, dokumentu-bilduma handiekin lan egiteko prestatua dagoena (Boldi eta Vigna, 2005).

Tresna honek aukera ematen du hitzen bilaketa soil bat baino gehiago egiteko. Esaterako, hitzak ordena jakin batean bila daitezke, edo hitzen arteko distantzia-mugaren bat erabiliz egin daitezke bilaketak. Horretaz gain, kontsulta batean indize bat baino gehiago konbinatzeko aukera ematen du.

Hainbat bilaketa-algoritmo ditu erabilgarri, besteak beste, *tf-idf* oinarrizko algoritmoa, edota, azken urteetan indar handia hartu duen BM25 algoritmoa (ikus 3.2.1 atala), eta honen aldaera berriena den BM25F algoritmoa.

5. kapituluko esperimentuak egiteko tresna hau erabili dugu.

3.4 Hizkuntzaren prozesamendurako teknikak

Sarreran aipatu bezala, hizkuntzek berezko duten sinonimia eta polisemia arazo izan daitezke IBa egiterakoan. Hori dela eta, proba eta esperimentu asko egin dira HPko teknikak IBko emaitzak hobetzeko erabil ote daitezkeen ikusteko.

Guk ere gure esperimentuetan hainbat HPrako teknika erabili ditugu eta horiexek azalduko ditugu jarraian.

3.4.1 Hitzen adiera-desanbiguazioa

Hara Lopez de Lacalle eta Agirrek (2010) nola definitzen duten hitzen adiera-desanbiguazioa (HAD):

> "Gure hizkuntza anbiguoa da. Hitz batek hainbat interpretazio ditu agertzen den testuinguruaren arabera, eta zein adiera hartzen duen asmatzea ez da lan erraza, nahiz eta guk era naturalean egin. Konputazio-metodoak erabiliz hitzen agerpenei adiera egokia ematea hitzaren adiera-desanbiguazioa (HAD) deritzo."

Hobeto ulertzeko, beraiek jartzen duten adibidea ere hona ekarriko dugu. Honako bi esaldi hauek emanik

(1) Parkeko *bankuan* eseri nintzen egunkaria irakurtzera.

(2) Dirua *bankuan* gorde dut.

eta *banku* hitzerako honako definizioak izanik Euskal WordNet ontologian (Agirre *et al.*, 2006)

banku-1: Eserleku luzea, bizkarduna nahiz bizkarrik gabea, hainbat lagun batera esertzeko aukera ematen duena.

banku-2: Bezeroen diru-gordailuak onartu eta kreditu-eragiketak egiten dituen enpresa publiko edo pribatua.

adibideko lehen esaldian HAD prozesu automatikoak *banku* hitzaren lehen adiera (*banku-1*) aukeratu beharko luke, eta bigarrenean *banku-2*.

Aipatu dugu dagoeneko IBko bi arazo nagusi polisemia eta sinonimia direla. HADak bai polisemia bai sinonimia arazoak automatikoki gainditzeko balio dezake.

4. kapituluko esperimentuetan erabilitako datu-multzoa, Robust izenarekin ezagutuko duguna, hitzen adierekin automatikoki etiketatua dago. 3.5.1 atalean daude datu-multzo honen inguruko xehetasun gehiago.

3.4.2 Ahaidetasun semantikoa UKB tresnaren bitartez

UKB tresnak grafo-algoritmoak erabiltzen ditu, eta, ezagutza-base lexikal batean oinarrituz, hitzen adiera-desanbiguaziorako eta antzekotasun/ahaidetasun[6] lexikala estimatzeko balio du (Agirre *et al.*, 2010c).

Grafo-algoritmo honen oinarrian *PageRank pertsonalizatua* deituriko algoritmoa dago. Eta, aldi berean, algoritmo hau PageRank algoritmo ezagunaren aldaera bat besterik ez da. Hortaz, lehenbizi PageRank algoritmoa zertan datzan ikusiko dugu.

Bilatzaile baten bidez indexatutako web-orrien garrantziari balioa emateko sortu zuten algoritmo-sorta da PageRank (Brin eta Page, 1998). Azken finean, grafo bateko erpinak beren egitura-garrantzia erlatiboaren arabera

[6] Antzekotasun semantikoa eta ahaidetasun semantikoa askotan sinonimo moduan erabiltzen badira ere, ez dute gauza bera adierazten. Antzekotasun semantikoak (*semantic similarity* ingelesez) sinonimia (auto-beribil) eta hiperonimia/hiponimia (taxi-auto) erlazioak hartzen ditu bere baitan. Ahaidetasun semantikoa (ingelesezko *semantic relatedness*etik itzuli dugu), ordea, orokorragoa da, eta meronimia (esku-hatz), antonimia (hotz-bero) eta asoziazioa (arkatza-papera) bezalako erlazioak ere kontuan hartzen ditu.

sailkatzeko balio du. PageRank metodoaren ideia nagusia honakoa da: grafo bateko i erpinetik j erpinera ertz zuzendu bat baldin badago, $i$tik $j$rako boto bat sortuko da, eta, ondorioz, j erpinak sailkapenean gora egingo du. Gainera, zenbat eta garrantzitsuagoa izan i erpina, orduan eta indartsuagoak izango dira $i$tik ateratzen diren botoak. Beste modu batera ere irudika daiteke algoritmo hau: grafoaren gainean ausazko bidea eginez (denbora-tarte luze batean) i erpinean amaitzeko probabilitatea adierazten du i erpinaren amaierako sailkapenak.

Eman dezagun G grafoa dugula N erpinekin eta i erpinetik ateratzen den ertz kopurua d_i dela, G grafoaren gaineko PageRank bektorea (P) honela kalkulatuko da:

$$\mathbf{P} = cM\mathbf{P} + (1 - c)\mathbf{v} \tag{3.6}$$

non $N \times N$ tamainako trantsizio-matrizea den M, $M_{ji} = \frac{1}{d_i}$ den $i$tik $j$ra ertz bat baldin badago, eta 0 bestela; $N \times 1$ tamainako matrizea den v eta bere elementu guztien balioa $\frac{1}{N}$ den; eta c moteltze-faktorea (*damping factor*) den, 0 eta 1 arteko balio eskalar bat hartzen duena. P bektorea ekuazio hau behin eta berriz kalkulatuz lortzen da, harik eta atalase batera heltzen den arte, edota aurrez finkatutako iterazio kopurua exekutatu arte.

3.6 ekuazioko lehenengo batugaiak lehen aipatutako boto-sistema egikaritzen du. Bigarren batugaiak, berriz, ausazko bide horretan erpin batetik bestera salto egiteko probabilitatea irudikatzen du. Esan dugun moduan, v bektoreko elementu guztiek $\frac{1}{N}$ balioa hartzen dute. Hortaz, grafoko erpin guztientzat berdina da ausaz erpin horretara salto egiteko probabilitatea.

Haveliwala-k (2002), ordea, v bektorea ez-uniformea izatea eta grafoko erpin batzuei probabilitate altuagoa esleitzea, alegia, erpin batzuei garrantzi handiagoa ematea proposatu zuen. Horrela, ausazko salto horietan erpin horietarako joera handiagoa izatea lortzen da. Esaterako, probabilitate-masa osoa i erpinari esleitzen badiogu, ausazko jauzi guztiak i erpinera itzuliko dira, honen maila igoaz. Eta, ondorioz, i erpinaren ondoko beste erpinen pisua ere handituko da. Hau da, i erpinari $v$ren hasierako banaketan emandako garrantzia grafoan zehar zabaltzen da algoritmoaren iterazio jarraietan. Aldaera honi PageRank pertsonalizatua deitzen zaio, eta horixe da UKB tresnak oinarrian duen algoritmoa.

UKB ezagutza-base lexikal baten gaineko algoritmo honen inplementazio bat da. Software librea da[7] eta guk gure esperimentuetan balio lehenetsiekin

[7]Hemen eskura daiteke: http://ixa2.si.ehu.es/ukb/

erabili dugu (c moteltze-faktorearen balioa 0,85 eta algoritmoa 30 iterazioren ondoren amaitu egiten da).

UKBrekin erabili nahi den ezagutza-base lexikal hori grafo moduan errepresentatu behar da. Hori edozein ezagutza-base lexikalekin egin badaiteke ere, publikoki eskuragarri dagoen inplementazio horrekin batera WordNeten grafo-errepresentazioa dago erabiltzeko prest. Eta guk horixe bera erabili dugu gure esperimentuetan.

UKB hitzen adiera-desanbiguazioa egiteko erabil badaiteke ere, guk testu bat emanik, testu horrekin ahaidetasun handiena duten WordNeteko kontzeptuak lortzeko erabili dugu. Bi atazak oso modu antzekoan egiten dira, baina hemen, guri dagokiguna azalduko dugu jarraian. Hasteko, erabili nahi dugun ezagutza-base lexikalarekin grafo bat osatu behar da. Grafo horren erpinak WordNeteko kontzeptuak (zehatzago esanda *synset*ak) eta hitzak (*variant*ak) izango dira. WordNeteko erlazioak *synset*en arteko ertz ez-zuzendu moduan jarriko ditugu; eta hiztegiko hitzak *synset*ekin lotzen dira ertz zuzenduen bidez. Honetarako WordNet 3.0 bertsioa erabili dugu, hor aurkitzen diren erlazio guztiak (baita glosetako erlazioak ere) grafoan jarriaz, ezarpen horiekin lortu baitzituzten hitzen antzekotasunerako emaitzarik onenak Agirre *et al.*-ek (2009c). Behin grafoa izanda, landu nahi dugun testua prozesatuko dugu bertako hitzen lemak eta kategoria gramatikalak lortzeko. Ondoren, grafoa hasieratuko dugu testuan dauden grafoko hitzei probabilitate uniforme bat esleituz. Hau da, hitz horiei dagokien erpinen artean banatuko dugu probabilitatea, eta beste erpin guztiei 0 balioa jarriko diegu hasieran. Eta, azkenik, PageRank algoritmo pertsonalizatua egikarituko dugu prestatutako grafo horren gainean. Horrela, WordNeteko kontzeptuen gainean probabilitate-banaketa gauzatzen da. Hortaz, exekuzioa amaitzean kontzeptu batek duen probabilitatea zenbat eta altuagoa izan, orduan eta ahaidetasun handiagoa izango du kontzeptu horrek landu dugun testuarekin.

Hauxe da UKBren bitartez testu batekin ahaidetasun handiena duten kontzeptuak, eta bide batez, hitzak, lortzeko bidea. 5. eta 6. kapituluetan azalduko ditugun esperimentuetan erabiliko dugu tresna hau kontsulten eta dokumentuen hedapena egiteko. Kapitulu horietan prozesu hau jarraituz egindako hedapenen adibideak ikus daitezke.

3.4.3 WordNet

WordNet ingeleseko ezagutza-base lexikal bat da[8] (Fellbaum, 1998); alegia, ingeleseko hitz eta adierei buruzko informazioa duen lexikoi bat da. Izen, aditz, adjektibo eta adberbioak aurkitzen dira bertan *synset* delakoen arabera antolatuta. *Synset* (*synonym set*) bat sinonimo-multzo bat dela esan dezakegu, kontzeptu lexikal edo adiera bati dagokiona. *Synset* bat hainbat ale lexikal edo hitzek osatzen dute, eta hitz hauek elkarren artean sinonimoak dira. Ale lexikal hauei *variant* deitzen zaie. Adibidez, *car, auto* eta *automobile* hitzak *synset* bateko *variant*ak dira, kontzeptu bera adierazten baitute. *Synset* bakoitzak adierazten duen kontzeptu hori, normalean, glosa edo definizio baten bidez adierazten da. Horrela, honako hau da aurreko adibidearen glosa: *a motor vehicle with four wheels*.

Hortaz, WordNeteko erlazio semantiko garrantzitsuenetako bat sinonimia da; baina ez bakarra. Izan ere, *synset*ak elkarren artean sinonimia ez den beste erlazio semantiko batzuen bidez erlazionatuta edo lotuta daude. Erlazio semantiko horien artean, garrantzitsuenetako bat hiperonimia-hiponimia erlazioa da. Hiperonimia-hiponimia erlazioak *synset* orokorrenak *synset* zehatzagoekin lotzen ditu. Lehengo adibidearekin jarraituz, {*car, auto, automobile*} *synset*aren hiperonimo bat *vehicle* da, eta *taxi*, berriz, *synset* horren hiponimo bat da.

Oraindik ere WordNet garatzen dihardute, eta momentu honetako (lan hau idatzi denekoa) bertsioa 3.0 da. Bertsio honetan 117.659 *synset* daude: 82.115 izen, 13.767 aditz, 18.156 adjektibo eta 3.621 adberbio[9]. Gure esperimentuetan, bestelakorik esaten ez bada behintzat, azken bertsio hau erabili dugu.

Gainera, beste hainbat hizkuntzatarako ere garatu dituzte WordNetak, hala nola, nederlanderako, italierarako, gaztelaniarako, alemanerako, frantseserako, txekierarako, estonierarako eta euskararako (Vossen, 1998; Pociello, 2008; Pociello *et al.*, 2010). Kontzeptu bera adierazten duten hizkuntza desberdinetako *synset*ek kode bera dute —edo hala ez bada, *synset* horien arteko mapaketa eskuragarri dago. Eleaniztasun hau dela eta, hitzen adieradesanbiguaziorako ez ezik, WordNet oso erabilia izan da HPko beste hainbat arlotan ere, hala nola, itzulpen automatikoan, galdera-erantzun sistemetan, informazio-erauzketan eta informazioaren berreskurapenean.

[8] http://wordnet.princeton.edu/

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

Guk kontsulta edota dokumentu batekin erlazionatutako kontzeptuak eta kontzeptu hauek adierazteko erabiltzen ditugun hitzak lortzeko erabili dugu WordNet. Hain zuzen ere, hitz horiek (*variant*ak) erabiliko ditugu kontsulta eta dokumentuen hedapena egiteko.

3.4.4 SemCor

SemCor semantikoki etiketatuako corpus bat da (Miller *et al.*, 1993). Hain zuzen ere, Brown Corpusaren zati bat, 200.000 hitz inguru dituena. Testu horietako izen, adjektibo, aditz eta adberbioak kategoria gramatikalekin automatikoki etiketatuta egoteaz gain, WordNeteko adierekin eskuz etiketatuta daude. Hasiera batean WordNeten 1.6 bertsioko adierekin etiketatu bazen ere, gerora bertsio berriagoetako adieretara egokitua izan da. WordNet sortu zuen talde berdinak landu du corpus hau.

Guk corpus hau erabili dugu ondorengo kapituluetan azalduko ditugun esperimentuetan zenbaitetan komeni zaigulako, kontzeptu jakin bat izanik, bere adiera usuenak zeintzuk diren jakitea. Horretarako, kontzeptu (*synset*) jakin hori adierazteko corpus honetan adiera bakoitzak dituen agerpen-maiztasunak erabili ditugu. *Variant* usuenak *synset* horrentzat batez bestekoa baino maiztasun handiagoa duten *variant*ak direla esango dugu.

3.4.5 Bestelakoak

Orain arte ikusitako HP teknika edo baliabideak semantika arlokoak dira. Badira, ordea, beste HP teknika orokorrago batzuk gaur egungo IB sistema gehienek erabiltzen dituztenak kontsulta eta dokumentuen aurreprozesuan, hala nola, tokenizazioa, *stopword*en ezabatzea, *stemming* edo lematizazioa eta hitz anitzeko unitate lexikalen trataera.

Tokenizazioa

Kontsulta edo dokumentu baten aurreprozesuan lehenik eta behin tokenizazioa egin behar da. Tokenizazioa deituriko prozesuan, indizea osatzeko eta oro har berreskurapen-prozesuan erabiliko diren unitateak izango diren token edo terminoak identifikatu behar dira testua osatzen duten karaktere segidan. Hasiera batean, terminoen arteko banatzaile moduan zuriuneak hartuz gauzatu daitekeela pentsa liteke, baina, hori baino prozesu konplexuagoa da.

Izan ere, besteak beste, digitu, marratxo, bestelako puntuazio-marka eta letra larriz idatzitakoen trataera nolakoa izango den argitu beharra dago (Fox, 1992).

Gure kasuan, esperimentu bakoitzean darabilgun IB sistemak eskaintzen duen tokenizatzailea erabiliko dugu, eta horiek eskaintzen dituzten aukeraz baliatuz, letra xehez ipiniko ditugu hitz guztiak.

*Stopword*en ezabatzea

2. kapituluan termino-haztapenaz jardun garenean esan dugun moduan, bilduman oso ohikoak diren terminoek ez dute ekarpenik egiten dokumentuak baztertzerako garaian, eta, hortaz, alferrikakoak dira berreskurapen-kontuetarako (Fox, 1992). Hitz hauek *stopword* moduan ezagutzen dira, eta horien adibide dira, esaterako, artikuluak, preposizioak eta juntagailuak. Aukera bat hitz hauek baztertzea da, indizetik eta berreskurapen-prozesuetatik kanpo utziaz. Gainera, hitz hauen baztertzearekin indizearen tamaina dezente txikitzen da. *Stopword*-zerrenda erabiltzea erabakitzen bada, zerrenda horretan zein hitz sartu erabaki behar da. Aukeretako bat bildumako hitz usuenak sartzea izango da. Baina, gerta daiteke hitz usuenak garrantzitsuegiak izatea baztertzeko, ingeleseko literaturan hitzik erabilienetakoak, adibidez, *time*, *war*, *home*, *life*, *water* eta *world* baitira (Fox, 1992). Hasieratik *stopword* hitzen zerrenda luzea erabiltzeko joera nagusitu izan da (200-300 hitzetakoa), baina azkenaldian zerrenda txikiak edo, web-bilatzaileen kasuan, horrelako zerrendarik ez erabiltzeko joera nagusitu da (Manning *et al.*, 2009). Izan ere, ohiko *stopword*-zerrendak erabiliz gero, kontsulta-mota jakin batzuk kaltetuak suertatzen dira. Esaterako, webean oso maiz egiten dira abestien izenburu edo bertso ezagunen bilaketak, eta horietan agertzen diren hitz asko *stopword*-zerrendan jarri ohi direnak izaten dira ("*To be or not to be*", "*Let it be*", "*I don't want to be*"...).

Gure esperimentuetako kontsultak tipologia horretakoak ez direnez, *stopword*ak baztertu ditugu kontsulta eta dokumentuetatik. Zerrenda orokor bera erabili dugu esperimentu eta datu-multzo guztietarako, hain zuzen ere, A eranskinean ikus daitekeena.

Stemming edo lematizazioa

Hizkuntza gehienetan hitz edo eratorpen morfologikoak aurki daitezke. HP ataza askotarako, eta baita IB atazatarako ere, komenigarria da tokenen

normalizazio morfologikoa egitea, hitz bat eta haren eratorpen morfologiko guztiak token beratzat hartzeko. Hori eginez gero, IB ataza batean aukera gehiago izango ditugu kontsultako hitzak dokumentuetako hitzekin parekatzeko, dokumentuetan kontsultako hitzen aldaera morfologikoak agertuta ere, topatuko ditugulako. Normalizazio morfologikoa egiteko oso erabilia den prozesuetako bat ingelesez *stemming* moduan ezagutzen dena da (Frakes, 1992). Prozesu hau egokia da morfologia sinpleko hizkuntzetarako, ingeleserako, adibidez, erregela finko batzuen bidez ohikoenak diren atzizkiak kentzen baitira. Baina, morfologia konplexuagoa duten hizkuntzetarako, gaztelaniarako edo euskararako, esaterako, egokiagoa da lematizatzaileak erabiltzea. Aplikazio hauek hitz bakoitzaren lema edo erroa lortzen dute, horretarako baliabide edo prozesu linguistikoak erabiliz, adibidez, hitzen analisi morfologikoa eginez eta hiztegi bat erabiliz.

Gure esperimentuetan ingeleseko testuak erabili ditugunean hizkuntza horretarako erabiliena den Porter *stemming*erako algoritmoa (Porter, 1980) erabili dugu. Algoritmo honek, adibidez, *advance*, *advanced*, *advancement*, *advancements*, *advances* eta *advancing* ingeleseko hitzak guztiak *advanc* bilakatzen ditu, eta *imagination*, *imaginations*, *imagine*, *imagining* eta *imaginings* hitzak *imagin* erroaz ordezkatzen ditu.

Hitz anitzeko unitate lexikalen trataera

Sag *et al.*-ek (2002) diotenaren arabera, 1.7 bertsioko WordNeteko sarreren % 41 hitz anitzeko unitateak (*multiword expressions*) dira; horien adibide dira honako hauek: *academic year*, *acquired immune deficiency syndrome*, *car insurance*, *carbon dioxide*, *even so*, *family tree*, *high school*, *Homo sapiens*, *human right*, *military service*, *mineral water*, *Mount Everest*, *move on*, *move up*, *neurotic depression*, *New York*, *Newton's law of gravitation*, *Nile River*, *nursery school*, *object-oriented programing language*, *operating system*, *Parkinson's disease*, *pick up*, *post meridiem*, *United Kingdom*, *vacuum cleaner*, *Valentine's Day*, *vertebral column*, *World War II* eta *World Wide Web*. Adibide hauetan ikusten den moduan, kasu batzuetan, horietako espresio osoaren esanahiak ez du zerikusirik espresioa osatzen duten osagaien esanahiarekin, edo beste modu batera esanda, esanahi berezi bat hartzen dute. Hori dela eta, HPko ataza askotan trataera berezi bat ematen zaie hitz anitzeko unitateei.

Gure esperimentuetan WordNet erabiltzen dugunez eta ikusirik WordNeten horrelakoen kopurua altua dela, esperimentu batzuetan horrelako es-

presioak ez ditugu banatzen, alegia, hitz anitzeko unitateak osorik indexatzen ditugu. Horretarako, dokumentuetako hitz-segidak WordNeteko hiztegiarekin parekatzen ditugu hauek identifikatzeko.

3.5 Datu-multzoak

Tesi-lan honetan erabili ditugun datu-multzoak banan-banan aurkeztuko ditugu jarraian. Datu-multzo horien hainbat datu (dokumentu eta kontsulta kopuruak eta hauen luzerak hitzetan neurtuak) jarri ditugu 3.1 taulan. Dokumentu kopuruak edo luzerak begiratzen baditugu, ikusten da izaera desberdineko datu-multzoak direla.

Datu-multzo guztietan bi kontsulta-bilduma dauzkagu: entrenamendurako eta testerako bildumak. Izan ere, guk erabili ditugun algoritmo guztiek dituzte parametroak, eta ahalik eta emaitzarik onenak lortzeko, hobe da parametro hauek doitzea. Baina, ez da komeni ebaluaziorako erabili nahi den bilduma erabiltzea parametro horiek doitzeko (Manning *et al.*, 2009):

> "Many systems contain various weights (often known as parameters) that can be adjusted to tune system performance. It is wrong to report results on a test collection which were obtained by tuning these parameters to maximize performance on that collection. That is because such tuning overstates the expected performance of the system, because the weights will be set to maximize performance on one particular set of queries rather than for a random sample of queries. In such cases, the correct procedure is to have one or more development test collections, and to tune the parameters on the development test collection. The tester then runs the system with those weights on the test collection and reports the results on that collection as an unbiased estimate of performance."

Guk ere, horixe bera egin dugu parametroak doitu ditugun kasu guztietan: entrenamendurako bildumaren gainean doitu parametroak, eta, ondoren, parametro horiek erabili testerako bilduma ebaluatzeko[10]. Azken ebaluazio horien emaitzak dira txosten honetan zehar aurkeztuko ditugun emaitza guztiak (besterik esaten ez bada behintzat).

[10]Parametro-doitzearen inguruko xehetasun gehiago 3.6 atalean emango ditugu.

ezaugarria	Robust	Yahoo!	ResPubliQA
dokumentu kopurua	166.754	89.610	1.379.011
dokumentu luzera (hitzak)	532	104	20
entrenamendu-kontsulta kopurua	150	1.000	100
test-kontsulta kopurua	160	30.000	500
test-kontsulta luzera (hitzak)	8,6	11,7	12,2

3.1 taula – Dokumentu kopurua, dokumentuen batez besteko luzera, entrenamendurako eta testerako kontsulta kopurua eta kontsulten batez besteko luzera datu-multzo bakoitzerako.

3.5.1 Robust-WSD

Datu-multzoetako bat *Cross-Language Evaluation Forum*eko (CLEF)[11] *Robust-WSD* atazetan erabilitakoa da (Agirre *et al.*, 2009a, 2010d). Ataza honen helburua HADak IB elebakar eta eleaniztunari zenbateko ekarpena egin diezaiokeen aztertzea zenez, datu-multzo honetako dokumentu eta gaiak hitzen adierekin etiketatuta daude. Ingeleseko datu-multzoaren etiketatzea artearen egoerako bi sistemek egin zuten (Agirre eta Lopez de Lacalle, 2007; Chan *et al.*, 2007), eta hortaz, datu-multzo honen bi bertsio etiketatu daude, UBC eta NUS moduan izendatuko ditugunak. Sistema hauek hitz bakoitzaren adierak WordNet 1.6 bertsioko *synset*ekin lotzen dituzte, lotura bakoitzari pisu bat esleituz, non pisurik handiena duen adiera den, sistema horren ustez, hitz horrentzako adierarik egokiena. Gaztelaniazko datu-multzoko hitz bakoitza bere lehen adierarekin etiketatuta dago. 3.1b irudian ingelesezko datu-multzo honetako desanbiguatutako gai bat ikus daiteke.

Dokumentuak LA Times 94 eta Glasgow Herald 95 egunkarietako berriak dira. Gaiek (ingelesez *topic* esaten zaie) informazio-beharra adierazteko hiru zati dituzte (ikus 3.1a irudiko adibidea): titulu motz bat (*title*); esaldi bakarreko deskribapen bat (*description)*; eta narratiba konplexuago bat (*narrative*), adierazgarritasunaren ebaluaziorako azalpenak emango dituena. Besterik esan ezean, gure esperimentuetan titulu eta deskribapen atalak erabili ditugu. Gai guztiak bi multzotan banatuta daude: entrenamendurako bilduma eta testerako bilduma. Gaiak ingelesez eta gaztelaniaz daude, eta dokumentuak, berriz, ingelesez bakarrik.

Etiketatutako datu-multzo hau erabili dugu 4. kapituluko esperimentue-

[11] http://www.clef-campaign.org/

ENtitle: Letter Bomb for Kiesbauer
ENdesc: Find information on the explosion of a letter bomb in the studio of the TV channel PRO7 presenter Arabella Kiesbauer.
ENnarr: A letter bomb from right-wing radicals sent to the black TV personality Arabella Kiesbauer exploded in a studio of the TV channel PRO7 on June 9th, 1995. An assistant was injured. All reports on the explosion and police inquiries after the event are relevant. Other reports on letter bomb attacks are of no interest.

(a) Gaia osorik, desanbiguatu gabe.

```
<ENtitle>
 <TERM ID="10.2452/141-WSD-AH-1" LEMA="letter" POS="NNP">
    <WF>Letter</WF>
    <SYNSET SCORE="0" CODE="05115901-n"/>
    <SYNSET SCORE="0" CODE="05362432-n"/>
    <SYNSET SCORE="0" CODE="05029514-n"/>
    <SYNSET SCORE="1" CODE="04968965-n"/>
  </TERM>
  <TERM ID="10.2452/141-WSD-AH-2" LEMA="bomb" POS="NNP">
    <WF>Bomb</WF>
    <SYNSET SCORE="0.888888888888889" CODE="02310834-n"/>
    <SYNSET SCORE="0" CODE="05484679-n"/>
    <SYNSET SCORE="0.111111111111111" CODE="02311368-n"/>
  </TERM>
  <TERM ID="10.2452/141-WSD-AH-3" LEMA="for" POS="IN">
    <WF>for</WF>
  </TERM>
  ...
</ENtitle>
<ENdesc>
  <TERM ID="10.2452/141-WSD-AH-5" LEMA="find" POS="VBP">
    <WF>Find</WF>
    <SYNSET SCORE="0" CODE="00658116-v"/>
    ...
  </TERM>
  ...
</ENdesc>
<ENnarr>
...
</ENnarr>
```

(b) Gaiaren zati bat desanbiguatuta

3.1 irudia – Robust-WSD datu-multzoko ingelesezko gai baten adibidea (10.2452/141-WSD-AH gaia).

Q: How do you cook an apple pie?

D: There are many good recipes for apple pies but there are also some important things to remember that are usually not in the recipe. That is you should make sure the bottom of the crust will bake as well and not remain soggy. To do this, coat the inside of the crust with butter before adding the filling and place the baking dish on a dark metal pan so the bottom will get more heat.

3.2 irudia – Yahoo! datu-multzoko galdera (Q) eta dokumentuaren (D) adibide bat (1005121203620 identifikadorea duena).

tan. Datu-multzo hau etiketatu gabe ere eskuragarri jarri zuten, eta etiketatu gabeko hori erabili dugu 5. eta 6. kapituluetako esperimentuetan.

3.5.2 Yahoo!

Yahoo! deituko diogun datu-multzoa *Yahoo! Answers* webgunearen iraulketa baten azpimultzo bat da[12]. Webgune honetan edonork galderak idatz ditzake, edota beste norbaitek egindako galderak erantzun. Datu-multzo hau osatzeko galdera guztien artetik ezaugarri linguistiko batzuk zituztenak aukeratu zituzten, esaterako, "how {to|do|did|does|can|would|could|should}"patroia betetzen zutenak. Horretaz gain, kalitate urriko erantzunak ere baztertu zituzten (Surdeanu *et al.*, 2008). Galdera horietako bakoitzaren erantzunik onenak bilduz osatu genuen dokumentu-bilduma (galdera bakoitzeko dokumentu adierazgarri bakarra dago). Galdera-bilduma osoa bi zatitan banatu dugu, entrenamendurako eta testerako bildumak sortuz. Datu-multzo honen ezaugarririk bereizgarriena galderak ez direla esperimentuetarako sortuak izan da, alegia, datu-multzo erreal bat da. 3.2 irudian datu-multzo honetako galdera eta honi dagokion erantzunaren dokumentua ikus daitezke adibide modura.

5. eta 6. kapituluetako esperimentuetan erabili dugu datu-multzo hau.

3.5.3 ResPubliQA

Beste datu-multzo bat CLEF 2009ko *Multilingual Question Answering* atazako ResPubliQA ariketarako prestatu zutena da (Peñas *et al.*, 2009). Ataza

[12] *Yahoo! Webscope dataset*en bitartez eskura daiteke: http://webscope.sandbox.yahoo.com/ ("ydata-yanswers-manner-questions-v1_0" datu-multzoa)

Q-eng: How fast does a tractor go?
Q-eusk: Zenbateko abiadura hartzen du traktoreak?
D: This Directive shall apply only to tractors defined in paragraph 1 which are fitted with pneumatic tyres and which have two axles and a maximum design speed between 6 and 25 kilometres per hour.

3.3 irudia – ResPubliQA datu-multzoko galdera ingelesez eta euskaraz (Q-eng eta Q-eusk) eta dokumentu (D) baten adibidea (96. galdera eta jrc31977L0537/14 dokumentua).

honetan lengoaia naturalean egindako 500 galderentzako erantzuna zuten pasarteak berreskuratu behar ziren. Dokumentu-bilduma *JRC-Acquis Multilingual Parallel Corpus*aren azpimultzo bat da, gutxi gorabehera 21.426 dokumentuz osatutakoa[13]. Dokumentu hauek, ingelesez eta Europako beste hainbat hizkuntza ofizialetan ere badaude. Eta, hortaz, galderak hizkuntza horietan ere badaude. Euskarazko dokumentuak ez baldin badaude ere, antolatzaileek euskarazko galderak prestatu zituzten. Horrela, datu-multzo hau erabiliz euskara-ingelesa hizkuntza arteko IBa egin daiteke. Hizkuntza-pare horrekin esperimentu batzuk egin baditugu ere, proba nagusiak ingelesarekin bakarrik (IB elebakarra) egin ditugu. 3.3 irudian datu-multzo honetako galdera (ingelesez eta euskaraz) eta honi dagokion erantzunaren dokumentua (ingelesez) ikus daitezke adibide modura.

Atazaren antolatzaileek eskuragarri jartzen duten datu-multzo honetan galdera eta dokumentuekin batera, adierazgarritasun-epaiak ere badaude. Galdera bakoitzeko adierazgarritasun-epai bakarra dago, hots, pasarte adierazgarri bakarra. Jakin badakigu, ordea, dokumentu-bilduma honetan galdera batzuentzat hainbat pasarte adierazgarri daudela. Hala ere, guk antolatzaileek banatutako adierazgarritasun-epai horiek erabili ditugu gure esperimentuak ebaluatzeko. Horretaz gain, entrenamendurako ingelesezko beste 100 galdera (eta hauei dagozkien adierazgarritasun-epaiak) prestatu eta eskuragarri jarri zituzten.

Hurrengo kapituluetan ikusiko dugu, besteak beste, dokumentuen hedapena egiten dugula. Dokumentu-bilduma honetan berreskuratu behar den unitatea pasartea denez, dokumentuak pasartetan zatitu ditugu. Hortaz, kasu honetan, pasarte bakoitzaren hedapena egingo dugu, baina, 10 hitz baino gehiago dituztenena bakarrik, ondorengo arrazoiak direla eta: alde batetik,

[13] 3.1 taulan datu-multzo honen datuetan agertzen dena pasarte kopurua da, hori baita indexatu dugun unitatea.

pasarte asko oso laburrak dira eta ez dute atazarako informazio baliagarririk (esaterako, "*Article 2*", "*Having regard to the proposal from the Commission*" edo "*HAS ADOPTED THIS REGULATION*"); bestetik, horrela, konputazio kostua asko murriztu dugu.

5. eta 6. kapituluetako esperimentuetan erabili dugu datu-multzo hau.

3.6 Parametro-doitzea

Hurrengo kapituluetan azalduko ditugun esperimentuetan erabilitako algoritmo eta sistemek hainbat parametro aske dituzte. 3.5 atalean esan bezala, parametro hauek datu-multzo bakoitzaren entrenamendurako bilduma erabiliz optimizatu ditugu. Esperimentu-multzo bakoitzean modu desberdin batean doitu ditugu parametroak.

Erabili dugun metodoetako bat Robertson eta Zaragoza-k (2009) deskribatutako *Promising Directions* izeneko metodoa da. Beste esperimentu batzuetan, aldiz, parametro bakoitzarentzat hainbat balio aukeratu eta *grid* bilaketa baten bidez aurkitu ditugu baliorik optimoenak. Esperimentuak azalduko ditugun kapituluetan zehaztuko dugu kasu horretan parametroak optimizatu ditugun edo ez, eta baiezko kasuan hauetako zein metodo erabili dugun.

3.7 Ebaluazioa

IB sistema bat ebaluatzeko bi ikuspuntu desberdin daude (Croft *et al.*, 2009): eraginkortasuna (*effectiveness*) eta errendimendua (*efficiency*). Eraginkortasunaren bidez, bilatzaileak informazio egokia bilatzeko duen gaitasuna neurtzen da. Eta, hitz gutxitan esanda, bilaketa horren azkartasuna neurtzen du errendimenduak. Alabaina, esan daiteke IBko ikerkuntzako lan gehienek eraginkortasuna hobetzen jartzen dutela arreta (Sanderson, 2010).

Ebaluazio hori egiteko testerako bilduma (*test collection*) erabili ohi da. Sanderson-en arabera, hauexek dira testerako bilduma baten osagai nagusiak:

- dokumentu-bilduma, non dokumentu bakoitzak *docid* identifikadore bat izango duen;
- gai edo kontsulta-sorta, non kontsulta bakoitzak *qid* identifikadore bat izango duen;
- adierazgarritasun-epaien sorta (*relevance judgments* edo *qrels* moduan

ezagutzen dena): *qid/docid* bikoteez osatutako zerrenda, kontsulta bakoitzeko dokumentuen adierazgarritasuna adieraziz. Adierazgarritasun hori balio bitarra izan ohi da: 1 balioa du adierazgarria bada, eta 0 bestela.

Horrelako testerako bilduma bat izanik, IB sisteman dokumentu-bilduma kargatu ondoren, kontsultak banan-banan egikaritzen dira. Sistemak, irteera moduan, kontsulta bakoitzerako itzulitako dokumentuak bilduko ditu fitxategi batean. Irteera honi *run* edo exekuzio deitzen zaio. Exekuzio hau adierazgarritasun-epaien sortarekin alderatu eta zenbateraino egin duen ondo neurtuz ebaluatuko dugu eraginkortasuna.

Ebaluazio honek 1950eko hamarkadan hasitako Cranfield-eko esperimentuetan du oinarria (Cleverdon, 1991). Hain zuzen ere, IBaren ebaluazioaren historia Cleverdon-ek egindako lanekin hasten dela esan ohi da. Ordutik gaurdaino, IB sistemen ebaluazioaz asko hitz egin da. Bereziki TREC[14], CLEF[15], FIRE[16] edota NTCIR[17] moduko nazioarteko biltzarrek garrantzi berezia hartu dute IB sistemen ebaluazioaren arloan azken urteetan. Biltzar hauen helburu nagusiak, besteak beste, IBrako test-bilduma berrerabilgarriak sortzea eta hauen ebaluazioa estandarizatzea dira. Guk CLEFeko hainbat atazatan parte hartu dugu, eta parte-hartze horietarako prestatutako esperimentuak dira, beste batzuen artean, hurrengo kapituluetan azalduko ditugunak. Gainera, TRECeko emaitzak ebaluatzeko sortu zuten *trec_ eval* aplikazioa, exekuzio bat eta adierazgarritasun-epaiak emanik, hainbat eraginkortasun-neurri kalkulatzen dituena. Aplikazio hau oso erabilia da IB komunitatean[18], eta guk ere horixe erabili dugu gure esperimentuak ebaluatzeko.

Ondorengo ataletan gure esperimentuen eraginkortasuna ebaluatzeko erabilitako neurriak zeintzuk diren ikusi, eta ebaluazioan erabilitako esanguratestaren (*significance test*) inguruko azalpen batzuk emango ditugu. Honen inguruko informazio gehiago nahi izanez gero, (Sanderson, 2010) lanean luze eta zabal hitz egiten da ebaluazioaren inguruan.

[14]Text REtrieval Conference: http://trec.nist.gov/

[15]Cross Language Evaluation Forum: http://www.clef-campaign.org/

[16]Forum for Information Retrieval Evaluation: http://www.isical.ac.in/~clia/

[17]NII Test Collection for IR Systems: http://research.nii.ac.jp/ntcir/index-en.html

[18]Hemen eskura daiteke: http://trec.nist.gov/trec_eval/

3.7.1 Eraginkortasun-neurriak

Doitasuna eta estaldura

Doitasuna (*precision*) eta estaldura (*recall*) dira IB sistema baten eraginkortasuna neurtzeko bi oinarrizko neurri. IBaren ebaluazioaren hastapenekin hasi zirenean definitu zituzten (Kent *et al.*, 1955; Cleverdon *et al.*, 1966), honela:

$$\text{doitasuna} = \frac{\text{berreskuratutako\_dokumentu\_adierazgarrien\_kopurua}}{\text{berreskuratutako\_dokumentuen\_kopurua}}$$

$$\text{estaldura} = \frac{\text{berreskuratutako\_dokumentu\_adierazgarrien\_kopurua}}{\text{dokumentu\_adierazgarrien\_kopurua}}$$

Hortaz, kontsulta batentzat berreskuratutako dokumentu-bildumatik adierazgarriak zenbat diren adierazten du doitasunak, eta berreskuratutako dokumentu adierazgarrien zatia zenbatekoa den adierazten du estaldurak.

Bi neurri hauek oso egokiak dira berreskuratutako dokumentuen zerrendan ordenarik ez bada zehazten dokumentuen artean. Baina, ikusi dugu ohikoena gaur egungo IB sistemek itzultzen duten dokumentuen zerrendan dokumentuen ranking bat egotea dela, alegia, lehen postuan dagoen dokumentua adierazgarriagoa dela bigarren postuan dagoena baino, eta hau hirugarrengoan dagoena baino, eta horrela jarraian datozen beste dokumentu guztiekin. Ebaluazioan ranking hori ere kontuan hartzeko beste neurri batzuk erabiltzen dira.

P@X

Horietako bat P@X moduan izendatuko duguna da (*precision at rank X*). Honek rankingeko X. posizioan doitasuna zenbatekoa den adierazten du:

$$P@X = \frac{\text{adierazgarriak}(X)}{X}$$

non $adierazgarriak(X)$ horrek rankingeko lehen X postuetan zenbat dokumentu adierazgarri dauden adierazten duen. Erabiltzailearen ikuspuntutik neurri hau interesgarria da. Izan ere, ohikoa da erabiltzaileak berreskuratutako zerrenda horretako hasieran dauden dokumentu gutxi batzuk begiratzea. Eta neurri honek, adibidez, lehen 5 edo 10 dokumentuak begiratuz

gero, horietatik adierazgarriak zenbat diren adierazten du (P@5 eta P@10, hurrenez hurren).

AP, MAP eta GMAP

Kasu batzuetan, ordea, lehen postuetako dokumentu gutxi batzuk aztertzea baino nahiago izango dugu ranking osoa ebaluatzea[19]. Batez besteko doitasunak edo *average precision* (AP) delakoak berreskuratutako dokumentu adierazgarri guztien posizioei dagokien doitasunaren batez bestekoa kalkulatzen du:

$$AP = \frac{\sum_{pos=1}^{N}(P@pos \times \text{adierazgarri}(pos))}{\text{dokumentu\_adierazgarrien\_kopurua}}$$

non N berreskuratutako dokumentu kopurua den, *pos* aztertzen ari garen rankingeko posizioa den eta $adierazgarri(pos)$ funtzio bitarrak 1 itzultzen duen *pos* posizioko dokumentua adierazgarria bada, eta 0 bestela. Neurri honek dokumentu adierazgarri guztien posizioak hartzen ditu kontuan, baina ebaluazioan eragin handia dute lehen postuetako dokumentu adierazgarriek. Hortaz, egokia da honakoa betetzen duten sistemak positiboki ebaluatzeko: ahalik eta dokumentu adierazgarri gehien berreskuratu eta gainera, dokumentu horiek lehen postuetan egotea.

AP neurriaren bidez kontsulta bakarra ebaluatzen dugu. Exekuzio oso bat (kontsulta bat baino gehiago) ebaluatzeko kontsulta bakoitzaren AParen batez bestekoa kalkulatzen duen *mean average precision* (MAP) erabili ohi da. Neurri hauxe da IB sistemen ebaluazioan erabiliena.

MAP neurriak badu arazo bat, ordea: ondoen gauzatzen diren galderetan egindako hobekuntzek txarto erantzundako galderen galerak estali egiten ditu. Beste modu batean esanda, galdera batzuk oso ondo eta beste batzuk oso gaizki erantzuten dituen sistema batek MAP altuagoa izan dezake galdera guztiak erdipurdi erantzuten dituen sistemak baino. Hori dela eta, ebaluazio batzuetan AParen batez besteko aritmetikoa erabili beharrean (hori da MAP), AParen batez besteko geometrikoa erabiltzen da (*geometric mean*

[19]Ranking osoa esatean, sistemak itzulitako dokumentu zerrenda osoa esan nahi da. IBaren arloan zabaldua dago oso zerrenda hori 1000 dokumentutakoa izatea gehienez. Guk ere gure esperimentuetan halaxe egin dugu

average precision edo GMAP) (Voorhees, 2005; Robertson, 2006):

$$GMAP = \sqrt[|Q|]{\prod_{|Q|} AP_{|Q|}}$$

non $\mid Q \mid$ ebaluatzen ari garen exekuzioko kontsulta kopurua den. Ikus dezakegun bezala, batez besteko doitasunen biderkadura egiten da, eta horrek galdera zailetan egindako hobekuntzen eragina nabarmentzen du emaitzan (galdera zailtzat hartzen dira AP baxua duten kontsultak). Demagun exekuzio bateko kontsulta baten APa 0,02tik 0,04ra hobetzen dela, baina beste kontsulta baten APa 0,4tik 0,38ra jaisten dela. Aldaketa horiek ez dute eraginik izango MAPean, batez besteko aritmetikoa berdina izango baita. Baina GMAPa altuagoa izango da, lehen kontsulta zail horretan egindako hobekuntza dela eta. Edo, adibidez, kontsulta batean 0,05etik 0,1erako aldaketak 0,25etik 0,5erako aldaketaren eragin berdina du GMAP neurrian, nahiz eta bigarren alde hori lehenengoa baino 5 aldiz handiagoa izan. Robust-WSD atazako ebaluazio ofizialetan neurri hau erabili zuten, eta guk ere, Robust datu-multzoarekin egindako esperimentu guztietan zehaztu dugu neurri hau.

MRR

Ataza batzuetan kontsulta bakoitzeko dokumentu adierazgarri bakarra egotea gerta liteke. Horrelako ataza baten eraginkortasun-neurriak dokumentu adierazgarriak lehen posizioetan berreskuratzeko lanak zenbateraino egiten dituen ondo ebaluatu beharko luke. Hori lehen ikusitako P@10arekin neurtu genezake, baina ez da oso zehatza. Esate baterako, P@10 neurriaren balioa berdina izango da dokumentu adierazgarri bakar hori lehen postuan berreskuratu edo hamargarren postuan berreskuratu. Hori dela eta, dokumentu adierazgarrien posizioarekiko sentikorragoa den MRR (*mean reciprocal rank*) neurria erabili ohi da horrelako kasuetan. *Reciprocal rank* deiturikoak lehen dokumentu adierazgarria zein posiziotan berreskuratzen duen hartzen du kontuan, hots, bigarren postuan berreskuratzen bada dokumentu adierazgarria, $1/2 = 0,5$ izango da bere balioa. Eta MRRak kontsulta guztien balio horren batez bestekoa kalkulatzen du.

$$MRR = \frac{1}{\mid Q \mid} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{pos}_{Q_i}}$$

non $\mid Q \mid$ ebaluatzen ari garen datu-multzoko kontsulta kopurua den eta pos_{Q_i} balioak Q_i kontsultarentzat berreskuratutako dokumentu guztien artean lehenengo dokumentu adierazgarria zein posiziotan dagoen adierazten duen.

ResPubliQA eta Yahoo! datu-multzoetan kontsulta bakoitzeko dokumentu adierazgarri bakarra dagoenez, neurri hau erabiliko dugu ebaluazio-neurri nagusi moduan.

c@1

CLEFeko ResPubliQA atazaren (ikus 3.5.3 atala) ebaluazioan honako ideia nagusi hau zuten: galdera bat erantzun gabe uztea hobea da gaizki erantzutea baino. Alegia, erantzun txarrak gutxitzeko galdera batzuk erantzun gabe uzten zituzten sistemak saritu nahi zituzten ebaluazioan, baina noski, beti ere ahalik eta galdera gehien ondo erantzunaz. Galdera bakoitzeko erantzun bakarra eman behar zen, eta, hortaz, ebaluatzaileek galdera batentzako hiru ebaluazio posible zituzten: (i) erantzun zuzeneko galdera, (ii) gaizki erantzundako galdera, eta (iii) erantzun gabe utzitako galdera. Eta horiek kontuan hartuz, neurri honen arabera ebaluatzen zuten exekuzio oso bat:

$$c@1 = \frac{1}{\mid Q \mid}\left(\mid Q \mid_Z + \mid Q \mid_G \frac{\mid Q \mid_Z}{\mid Q \mid}\right)$$

non $\mid Q \mid$ exekuzioko galdera guztien kopurua den, $\mid Q \mid_Z$ zuzen erantzundako galdera kopurua den, eta $\mid Q \mid_G$ erantzun gabeko galdera kopurua den. Neurri honen arabera, sistema batek galdera guztiak erantzuten baditu, $\mid Q \mid_G = 0$ izango denez, galdera guztietatik ondo erantzundako galdera kopurua zenbatekoa den neurtuko da ($c@1 = \frac{|Q|_Z}{|Q|}$). Eta, aldiz, galdera guztiak erantzun gabe uzten baditu, $c@1 = 0$ neurria izango du, $\mid Q \mid_Z = 0$ izango baita.

ResPubliQA 2010 atazako parte-hartzeari buruz hitz egitean, neurri hau ere komentatuko dugu (ikus 5.5.8 atala).

3.7.2 Esangura-testak

Aurreko atalean ikusi dugu IBko sistema bat ebaluatzeko erabil daitezkeen neurriak zeintzuk diren. Baina, nola konparatuko ditugu bi sistema? Noiz esango dugu sistema bat bestea baino hobea dela? Aukera bat izan zitekeen sistema batek beste sistema batek baino MAP altuagoa lortzen badu,

lehenengo sistema hori bigarrena baino hobea dela esatea. Baina hori ez litzateke guztiz zuzena izango. Izan ere, baliteke kontsulta, dokumentu eta adierazgarritasun-epai zehatz horiek emanik, zoriz gertatzea lehen sistema horrek bigarrenak baino hobeto egitea. Hortaz, beharrezkoa da esangura-testak (*significance tests*) erabiltzea bata bestea baino hobea den (edo hala ez den) jakiteko, test hauen bitartez jakingo baitugu bi sistemen ebaluazioen artean dauden diferentziak estatistikoki esanguratsuak diren edo ez (beti ere huts-egite probabilitate bat kontuan izanik).

IBko hasierako lanetan ebaluazioari garrantzi handia ematen bazioten ere, esangura-testak ez ziren oso erabiliak. Hull-ek (1993) hauek erabiltzea beharrezkoa zela esan zuen, eta azkenaldian, hauen erabilera handituz doa. Smucker *et al.*-ek (2007) IBrako egokiak izan daitezkeen esangura-testak aztertzen dituzte, eta *Paired Randomization Test* erabiltzea gomendatzen dute. Guk hurrengo kapituluetako esperimentuen ebaluazioan horixe bera erabili dugu.

4

Adiera-desanbiguazioa eta hizkuntza-ereduetan oinarritutako IBa

Kapitulu honetan hitzen adiera-desanbiguazioa erabiliz IB sistemaren eraginkortasuna hobetzeko helburuarekin egindako esperimentuak azalduko ditugu. Horretarako, hitzen adiera-desanbiguazioa eta ezagutza-base lexikal bat (WordNet) baliatuz, kontsulta eta dokumentuei sinonimoak gehituz hedapena egitea proposatzen dugu, ondoren, hedapen hori hizkuntza-ereduetan oinarritutako IB sistema batean txertatzeko. Esperimentu hauek *Robust-WSD Task @ CLEF 2008* atazan parte hartzeko egin genituenak dira; ataza elebakarreko (ingelesa) eta hizkuntza arteko atazako (gaztelania-ingelesa) esperimentuak dira, hain zuzen ere. Hori dela eta, ataza horretako antolatzaileek prestatutako datu-multzoa erabili dugu, zeina adiera-desanbiguatzaile baten bidez etiketatuta dagoen.

4.1 Aurrekariak

Sarreran aipatu bezala, kapitulu honetan landuko ditugun esperimentuen motibazioa *Robust-WSD Task @ CLEF 2008* (Agirre *et al.*, 2009a) atazatik dator, bertan parte hartzeko helburuarekin egindako esperimentuak baitira jarraian azalduko ditugunak.

Ataza honen aurrekaria beste ataza bat izan zen, *SemEval-2007 Task 01: Evaluating WSD on Cross-Language Information Retrieval* (Agirre *et al.*, 2008). Ataza honen helburu nagusia HADa aplikazio baten baitan ebaluatzea zen, bi helburu zehatzago zituelarik: alde batetik, zein HAD estrategia

zen onena ikustea, eta bestetik, HADa IB sistema batean erabiltzea onuragarria zela erakustea. Horretarako, antolatzaileek ebaluazioan erabiliko ziren datu-multzoaren hedapen-prozesua eta IB sistema prestatu zituzten. Parte-hartzaileek datu-multzo jakin bat desanbiguatu beharra zuten beraien HAD sistemarekin. Ataza honetako saiakeretan HAD informazioa erabiliz oinarrilerroko sistema hobetzerik lortu ez bazen ere, emaitza itxaropentsuak lortu zituzten.

Hori dela eta, ikerketa-lerro hau gehiago aztertzeko asmoz, aipatutako *Robust-WSD Task @ CLEF 2008* ataza antolatu zuten, oso antzeko ezaugarriekin. Helburua, berriz ere, adiera-desanbiguazioak IBan (elebakar nahiz hizkuntza artekoan) egin ditzakeen ekarpenak aztertzea zen. Horretarako, ataza honetan antolatzaileek parte-hartzaileei HAD sistema batekin etiketatutako datu-multzo bat jarri zien eskuragarri; zehatzago esanda, datu-multzo beraren bi aldaera zeuden, bi HAD sistemekin etiketatuta (datu-multzo honen inguruko datu gehiago 3.5.1 atalean aurki daitezke). Horrela parte-hartzaileek datu-multzo hori erabili beharra zuten, baina beraiek nahi zuten IB sistema erabiltzeko eta HAD informazio hori nahi zuten bezala ustiatzeko aukera zuten. Emaitzei dagokienez, ataza honetako parte-hartzaile batzuk HAD informazioa erabiliz emaitzak hobetzeko gai izan ziren ataza honetan; guk ere lortu genuen, emaitzen atalean ikusiko dugun moduan.

Aipatu, ataza honekin batera oso antzeko beste ataza bat antolatu zela galdera-erantzun sistemetarako ere, *QA-WSD @ CLEF 2008* ataza, hain zuzen ere (Forner *et al.*, 2009).

4.2 Hitzen adiera-desanbiguazioa testuaren hedapenerako

HADa IBrako baliagarria den edo ez ikusteko, HADetik eratorritako informazio hori dokumentu edota kontsultak hedatzeko erabil daiteke. Atal honetan hori gauzatzeko modu bat aurkeztuko dugu.

Horretarako, Robust-WSD datu-multzoko gai eta dokumentuak erabiliko ditugu[1]. Gai eta dokumentu hauek WordNeteko adierekin edo *synset*ekin etiketatuta daude eta etiketa bakoitzak pisu bat du esleitua, non pisurik handiena duen adiera den, HAD sistemaren ustez, hitz horrentzako adierarik egokiena. Desanbiguatutako gaztelaniazko gai-bilduma bat eta ingelesezko

[1]Gogoratu, *gai* deituko diegu ingelesez *topic* moduan ezagutzen diren adierazpenei.

gai eta dokumentuen bina bilduma daude (azken multzo hauek UBC eta NUS moduan izendatuak).

Desanbiguatutako datu-multzo horiek eta ingelesezko eta gaztelaniazko WordNetak erabiliz gai eta dokumentuen hedapenak egingo ditugu, honela: hitz bati lotutako *synset*ari dagozkion *variant*ak esleituko dizkiogu hitz horri. *Synset* bateko *variant*ak sinonimoak direnez, hitz horri bere sinonimoak gehitzen ari gara modu honetan. Adiera batek hizkuntza guztietako WordNetetan synset zenbaki bera du –edo mapaketa bidez erraz lor daiteke beste hizkuntza bateko *synset* zenbakia. Hortaz, hedapen hori hizkuntza batetik bestera ere gauzatu daiteke. Kasu hauetan, hitz bati esleituko dizkiogun hitzak *synset* horrek beste hizkuntzako WordNetean dituen *variant*ak izango dira, hots, hitz horren beste hizkuntzako itzulpenak izango dira. Hori izango da ingelesezko gai eta dokumentuak gaztelaniara eta gaztelaniazko gaiak ingelesera itzultzeko erabiliko dugun estrategia.

Oinarrizko hedapen-eredu horri jarraituz, hainbat hedapen eta itzulpen gauzatu ditugu. Jarraian, egindako hedapen eta itzulpen mota desberdinak zeintzuk diren ikusiko dugu, adibide batzuez lagundurik. Adibideak 4.1, 4.2 eta 4.3 irudietan ikus daitezke. Adibide horietako bakoitzean, lehenik, jatorrizko gai edo dokumentuak ikus daitezke, eta, ondoren, horiei dagozkien hedapen edo itzulpen adibideak. Jatorrizko gaietan (4.1a eta 4.3a) gako-hitzak **kolore honetan** markatu ditugu[2], bakarrik hitz horien hedapen edo itzulpena egiten dugulako. Hedapenen adibideetan lerro bakoitzean hitz bakoitzaren hedapenetik lortutako hitzak jarri ditugu. Hedapen edo itzulpen osoaren adibideetan adiera bakoitzetik lortzen diren sinonimo-multzoak / batez bereizi ditugu. Hedapenean gako-hitzetako bat ez bada azaltzen ezin izan dugulako hedatu izango da. Errepikatzen diren hitzak behin bakarrik jarri ditugu adibideetan.

Hona bada hedapen kasu desberdinen zerrenda:

– **Ingelesezko gai eta dokumentuen hedapen osoa**: hitz bakoitzaren adiera guztien sinonimo guztietara hedatzea. Kasu honetan ez da HADeko informazioa erabiltzen. 4.1b irudian ikus daiteke gai baten mota honetako hedapenaren adibide bat (modu berean egingo litzate-

[2]Gai hauetan ohiko *stopword* hitzak ezabatzeaz gain (ikus 3.4.5 atala), datu-multzo hauetan ohikoa den bezala, beste hainbat hitz ere ez ditugu kontuan hartzen, hain zuzen ere, ia gai guztietan errepikatzen direnak: ingelesezko gaietan, besteak beste, *find*, *describing*, *discussing*, *documents* eta *report*; gaztelaniazko gaietan *encontrar*, *describir*, *documentos*, *noticias*, eta *ejemplos*.

ENtitle: **Computer Mouse RSI**
ENdesc: Find documents that report on **computer mouse repetitive strain injuries** (**RSI**).

(a) Jatorrizko ingelesezko gaia.

computer → calculator, computer, estimator, figurer, reckoner / computer, data processor, electronic computer, information processing system
mouse → mouse / mouse
repetitive → insistent, repetitive / iterative, reiterative, repetitious, repetitive
strain → song, strain / breed, strain / strain / strain, straining, stress / strain, stress / mental, strain, nervous, strain, strain / air, line, melodic line, melodic phrase, melody, strain, tune / breed, stock, strain, variety / strain, tenor / nisus, pains, strain, striving / strain / form, strain, var., variant
injuries → accidental injury, injury / harm, hurt, injury, trauma / combat injury, injury, wound / injury

(b) Gaiaren hedapen osotik lortutako hitzak.

computer → computer, data processor, electronic computer, information processing system
mouse → mouse
repetitive → insistent, repetitive
strain → strain
injuries → harm, hurt, injury, trauma

(c) Gaiaren hedapen onena jarraituz lortutako hitzak (UBC bertsioa).

4.1 irudia – Robust-WSD datu-multzoko gaiaren (10.2452/064-AH) hedapenen adibideak.

ke dokumentu baten hedapena).

– **Ingelesezko gai eta dokumentuen hedapen onena**: hitz bakoitzaren pisurik handieneko adieraren sinonimo guztietara hedatzea. Hedapen honen bi bertsio izango ditugu, UBC eta NUS sistemen datuetan oinarritutakoak. 4.1c irudian ikus daiteke gai baten hedapen honen adibide bat (modu berean egiten da dokumentu baten hedapena).

Eta hauek dira egindako itzulpen desberdinak[3]:

– **Ingelesezko dokumentuen itzulpen osoa**: hitz bakoitzaren adiera guztien itzulpena ingelesetik gaztelaniara. 4.2b irudian ikus daiteke honelako itzulpen baten adibidea.

[3]Ingelesezko dokumentuak eta gaztelaniazko gaiak dira itzuliko ditugunak.

– **Ingelesezko dokumentuen itzulpen onena**: hitz bakoitzaren pisurik handieneko adieraren itzulpena ingelesetik gaztelaniara. Itzulpen honen bi bertsio izango ditugu, UBC eta NUS sistemen datuetan oinarritutakoak. 4.2c irudian ikus daiteke mota honetako itzulpen baten adibidea.

– **Gaztelaniazko gaien itzulpen onena**: hitz bakoitzaren lehenengo adieraren itzulpena gaztelaniatik ingelesera[4]. Hitz batek ez badu itzulpenik (WordNeten ez dagoelako hitz hori), jatorrizko gaztelaniazko hitz hori jarriko dugu itzulpenean. 4.3b irudian adibidea. *Síndrome RSI* (hitz anitzeko unitate lexikal moduan) eta *RSI* hitzak WordNeten ez direnez agertzen, hitz horiek berak jarriko ditugu itzulpenean.

Adibide horiei guztiei dagozkien XML fitxategiak B eranskinean aurki daitezke. Bertan, gai edo dokumentu bakoitzarentzat HAD sistemak itzulitako fitxategiak eta fitxategi horien gainean egindako hedapenaren XML fitxategiak jarri ditugu. Adibide hauetako hedapenetako hitz bakoitza nondik eta nola lortu den ikusteko aukera dago fitxategi horietan.

4.3 Desanbiguazioan oinarritutako hedapen-ereduak IB sistema baterako

Aurreko atalean proposatutako hedapen-eredu horiek erabiliz, hainbat gai- eta dokumentu-hedapenen multzo izango ditugu. Horiekin hainbat indize eta kontsulta-sorta sortuko ditugu, eta horiek konbinatuz, IBko hainbat exekuzio desberdin gauzatuko ditugu. Jarraian hedapen horiek IB sisteman nola txertatu ditugun azalduko dugu.

4.3.1 Desanbiguazioan oinarritutako dokumentu-hedapena IBrako

Behin aurreko atalean proposatutako eredu bakoitzaren bidez dokumentuen hedapena (edo itzulpena) burututa, hortik sortutako dokumentu-bilduma berri bakoitzarekin indize bat sortu dugu. Hortaz, indize horietako bat dokumentuetako jatorrizko hitzekin bakarrik osatuta egongo da. Besteetan, aldiz,

[4]Gaztelaniazko datu-multzo honetarako antolatzaileek zabaldu zuten bakarra lehenengo adieraz etiketatutakoa zen.

2 FIRMS ADOPT LABELS WARNING COMPUTER USERS ABOUT DANGER OF INJURY; TECHNOLOGY: COMPAQ, MICROSOFT ARE THE FIRST TO STATE THAT HARM COULD COME FROM KEYBOARD MISUSE OR TOO MUCH TYPING.

(a) Jatorrizko ingelesezko dokumentuaren zati bat.

labels → etiqueta / etiqueta, pegatina, pegata, etiqueta adhesiva / marca
warning → avisar, notificar / amonestar, prevenir, avisar
computer → ordenador, procesador
user → explotador / consumidor de drogas / usuario
danger → peligro, riesgo / peligro / peligro
injury → trauma, daño, contusión, herida, lesión, traumatismo / herida accidental / agravio / herida, herida de guerra
technology → ingeniería, tecnología / ciencia aplicada, tecnología, ingeniería
are → representar, ser / ser / trabajar / existir, haber / caracterizar, personificar, encarnar / estar, haber / costar, valer / acaecer, suceder, ocurrir / durar, vivir, existir / equivaler, significar, ser equivalente
first → primero, primera / primero, primera / primera base / comienzo, umbral, principio, inicio / primera
state → departamento de estado / estado federal, estado / estado, país, república, nación / estado / tierra, estado, país, nación / estado de la materia, estado físico
harm → daño, perjuicio / trauma, daño, contusión, herida, lesión, traumatismo / daño, perjuicio, mal, destrozo
keyboard → teclado
misuse → abuso, desaprovechamiento
much → mucho
typing → mecanografiar

(b) Dokumentuaren itzulpen osoa jarraituz lortutako hitzak.

labels → etiqueta
warning → avisar, notificar
computer → ordenador, procesador
user → usuario
danger → peligro, riesgo
injury → trauma, daño, contusión, herida, lesión, traumatismo
technology → ingeniería, tecnología
are → ser
first → primero, primera
harm → trauma, daño, contusión, herida, lesión, traumatismo
keyboard → keyboard
misuse → abuso, desaprovechamiento
much → mucho
typing → mecanografiar

(c) Dokumentuaren itzulpen onena jarraituz lortutako hitzak (UBC bertsioa).

4.2 irudia – Robust-WSD datu-multzoko ingelesezko dokumentuaren (LA081794-0225) itzulpenen adibideak.

EStitle: **Síndrome RSI** y **ratones** de **ordenador**
ESdesc: Encontrar documentos que informen sobre **RSI** ("**repetitive strain injuries**"o "**enfermedad** del **periodista**") **producidas** por el **uso** del **ratón** del **ordenador**.

(a) Jatorrizko gaztelaniazko gaia.

Síndrome RSI → Síndrome RSI
ratones → mouse
ordenador → computer, data processor, electronic computer, information processing system
RSI → RSI
repetitive → repetitive
strain → strain
injuries → abuse, blackguard, clapperclaw, shout
enfermedad → disease
periodista → newsman, newsperson, reporter
producidas → create, make, produce
uso → employment, exercise, usage, use, utilisation, utilization

(b) Gaztelaniazko gaiaren itzulpen onena jarraituz lortutako hitzak.

4.3 irudia – Robust-WSD datu-multzoko gaiaren (10.2452/064-AH) itzulpenaren adibidea.

jarraitu den ereduaren arabera, jatorrizko hitz horiez gain hedapenetik lortutako hitzak ere egongo dira indizean. Indize guztietatik, 3.4.5 atalean esan bezala, *stopword* hitzak ezabatu ditugu.

4.3.2 Desanbiguazioan oinarritutako kontsulta-hedapena IBrako

4.2 atalean esandako moduan gaien hedapena egin ostean, honako kontsulta-sorta hauek izango ditugu:

(i) Ingelesezko kontsultak, jatorrizko hitz eta hitz bakoitzaren sinonimo guztiekin.

(ii) Ingelesezko kontsultak, jatorrizko hitz eta hitz bakoitzaren adiera onenaren sinonimo guztiekin.

(iii) Gaztelaniazko kontsultetako hitz bakoitzaren lehenengo adieraren itzulpenekin osatutako ingelesezko kontsultak.

Ikus dezagun jarraian kontsulta horietan oinarrituta nola egiten den berreskurapena. Zerrendatutako horiek dira desanbiguazioan oinarritutako kontsulta-hedapenak, DQE (*Disambiguation based Query Expansion*) moduan izendatuko ditugunak. IBa egiteko, kontsulta-sorta bati hedapen hauetako bat aplikatu eta dokumentuak berreskuratuko ditugu. Arestian ikusitako (i) eta (ii) kasuetan, dokumentu batetik kontsulta hedatu osoa (Q_{DQE}) sortzeko probabilitatearen arabera sailkatuko ditugu dokumentuak, probabilitate hori honela kalkulatuko dugularik:

$$P_{DQE}(Q_{DQE} \mid \Theta_D) = P(Q \mid \Theta_D)^w P(Q' \mid \Theta_D)^{1-w} \tag{4.1}$$

non Θ_D D dokumentuaren hizkuntza-eredua den, Q jatorrizko kontsultaren hedapena Q' den eta w jatorrizko kontsultari esleitutako pisua den (0-1 tartean finkatu beharrekoa).

Jatorrizko kontsultari dagokion probabilitatearentzat *query likelihood*aren bidez egingo dugu estimazioa banaketa multinomialari jarraituz, honela:

$$P(Q \mid \Theta_D) = \prod_{i=1}^{|Q|} P(q_i \mid \Theta_D)^{\frac{1}{|Q|}} \tag{4.2}$$

non Q kontsultako terminoa den q_i eta $|Q|$ kontsulta horren luzera den (termino kopurua). *Dirichlet leuntze-teknika* (Zhai eta Lafferty, 2001a) jarraituz honako hau dugu:

$$P(q_i \mid \Theta_D) = \frac{\mathrm{tf}_{q_i D} + \mu \frac{\mathrm{tf}_{q_i C}}{|C|}}{|D| + \mu} \tag{4.3}$$

non $\mathrm{tf}_{q_i D}$ eta $\mathrm{tf}_{q_i C}$ D dokumentuko eta bilduma osoko q_i kontsulta-terminoaren maiztasunak diren, hurrenez hurren, eta μ leuntze-teknikaren parametro askea den.

Hedatutako kontsultaren *query likelihood* estimazioa ($P(Q' \mid \Theta_D)$) jatorrizko kontsultarenaren modu berdinean egiten da (4.2 ekuazioa jarraituz baita). Aldatzen dena termino bakoitzaren estimazioa da; 4.3 ekuazioa beharrean, honako hau izango dugu hedatuko dugun termino bakoitzerako:

$$P(q'_i \mid \Theta_D) = \frac{\mathrm{tf}_{q'_i D} + \mu \frac{\mathrm{tf}_{q'_i C}}{|C|}}{|D| + \mu} \tag{4.4}$$

non q'_i hitz-multzo bat den, hain zuzen ere, jatorrizko q_i terminoaren hedapenetik lortutako hitzez osatutako multzoa (adibideko irudietan komaz bereizirik ageri direnak). Alegia, hitz-multzo horretako hitzak elkarren artean

sinonimoak izango dira, eta horregatik sinonimo-multzo osoaren maiztasunak hartuko dira kontuan:

$$\mathrm{tf}_{q_i' D} = \sum_{i=1}^{|S|} \mathrm{tf}_{s_i D}$$

$$\mathrm{tf}_{q_i' C} = \sum_{i=1}^{|S|} \mathrm{tf}_{s_i C}$$

non jatorrizko kontsultako termino baten hedapeneko hitzen multzoa den S, alegia, s_i sinonimoen multzoa.

Itzulpenetik sortutako kontsulta-sortaren kasuan ere (aurreko zerrendako (iii) kasua), dokumentu batetik itzulitako kontsulta hori sortzeko probabilitatearen arabera sailkatuko dira dokumentuak. Alegia, *query likelihood* probabilitatearen arabera sailkatuko dira, eta, honetan ere, banaketa multinomiala eta Dirichlet leuntze-teknika jarraituz kalkulatuko da probabilitate hori, 4.2 eta 4.4 ekuazioetan zehaztu bezala (honetan ere termino baten itzulpenetik sortutako sinonimoak multzokatuta tratatuko dira).

Aipatu, q_i moduan adierazi dugun termino hori hitz anitzeko unitate lexikala izan daitekeela, WordNeteko hiztegia erabiliz kontsultan dauden hitz anitzeko unitate lexikalak identifikatzen baititugu. Horren adibide dira, besteak beste, *North American* (146 identifikadorea duen gaian agertzen da), *ozone layer* (148), *death penalty* (150), *CD burner* (169), *ice hockey* (171), *political party* (186), *black hole* (189), *alternative medicine* (251), *World War II* (274), *land mine* (274), *mercy killing* (277), *public transport* (278), *nuclear power* (298), *Mexico City* (327) eta *grand slam* (346). s_i moduan izendatu ditugun terminoak ere izan daitezke hitz anitzeko unitate lexikalak.

4.4 Esperimentazio-ingurunea

Kapitulu honetan egindako esperimentu nagusiak *Robust WSD Task @ CLEF 2008* atazan parte hartzeko eginak dira. Hortaz, esperimentazio-ingurunea ataza horren antolatzaileek zehaztutakoa da. Alde batetik, Robust-WSD moduan izendatu dugun datu-multzoa erabili behar genuen. Datu-multzo honetako gai nahiz dokumentuak hitzen adiera-desanbiguazioko informazioarekin aberastuta daude. Bi gai-bilduma genituen: bata entrenamendurakoa eta bestea, berriz, ataza honen lehiaketa-fasean eskuratu ahal izan genuen testerako gai-bilduma. Gainera, ingelesezko desanbiguazioa bi sistemek egin

zutenez, ingelesezko gai eta dokumentuen bina bertsio genituen (UBC eta NUS moduan izendatuko ditugunak). Gaztelaniaz gaiak bakarrik genituen. Bestetik, IB elebakar (gai eta dokumentuak ingelesez) edo hizkuntza artekoa (gaztelaniazko gaiak eta ingelesezko dokumentuak) egin genezakeen. Ataza honen helburua HADak IBari ekarpenik egin ote ziezaiokeen frogatzea zenez, parte-hartzaileek HAD informaziorik erabili gabeko oinarri-lerro bat eta HAD informazioa erabiliz egindako beste exekuzio bat, gutxienez, bidali behar zuten ataza honetara.

Guk esperimentu hauek burutzeko (indizeak sortzeko, hedapen-ereduak inplementatzeko, IB exekuzioak gauzatzeko) Indri bilatzailea erabili dugu (Strohman *et al.*, 2005) (begiratu 3.3.1 atalean tresna honen inguruko zehaztapenak).

Indizeak sortzerako garaian Krovetz *stemmer*a erabili dugu. HAD informazio horrekin batera hitz bakoitzaren kategoria gramatikala zehaztuta zegoenez, informazio hori aprobetxatu eta izen, adjektibo, aditz eta zenbakiak bakarrik indexatu ditugu.

Esan bezala, HAD informaziorik gabeko oinarri-lerro bat prestatu beharra genuen. Guk bi oinarri-lerro prestatu ditugu. Bata, Indri sisteman eredu lehenetsia den *query likelihood* hizkuntza-eredua erabiliz lortutakoa, eta, bestea, *pseudo-relevance feedback* (PRF) deitzen dena. Azken hau ere, Indri sisteman inplementatuta dago, eta Lavrenko-k proposatutako adierazgarritasun-ereduaren (ingelesezko *relevance model*) aldaera bat da (Lavrenko eta Croft, 2001), non berreskurapena egiteko erabiltzen den jatorrizko kontsultaren eta hedatutako kontsultaren konbinazio bat den. Bi oinarri-lerro hauetarako ere Dirichlet leuntze-teknika erabili dugu. PRFak datu-multzo askotan hobekuntzak lortzen dituenez (Buckley eta Sanderson, 2008), aurreko atalean azaldutako hedapen-ereduei ere, PRFa aplikatu diegu.

Erabaki hori hartzeko, eta testerako datu-multzoarekin hasi aurretik, entrenamenduko datu-multzoarekin hainbat esperimentu egin ditugu eta ikusi dugu gure kasuan ere PRFarekin emaitza hobeak lortzen genituela. Proba hauetan indize desberdinak kontsulta-sorta desberdinekin konbinatu ditugu eta MAP altuena lortzen zutenak hartu ditugu, testerako bildumarekin konbinazio horiek berak egin eta lehiaketarako exekuzio ofizial moduan aurkezteko.

Baina proba horietan ez ditugu parametroak optimizatu. Hortaz, Dirichlet leuntze-teknikaren μ parametroari (4.3 ekuazioa) balio lehenetsia esleitu diogu (Indriren inplementazioan balio lehenetsi hori 2500 da). PRF ereduak 3 parametro ditu: dokumentu eta termino kopurua (*fbDocs* eta *fbTerms*),

eta *fbOrigWeight* jatorrizko kontsultaren pisua. Hauei ere balio lehenetsiak eman dizkiegu: $fbDocs = 10$, $fbTerms = 50$ eta $fbOrigWeight = 0,5$. 4.1 ekuazioko w pisuari 0,6ko balioa eman diogu.

4.5 Emaitzak eta analisiak

Robust-WSD ataza elebakar eta hizkuntza artekoan hartu genuen parte. Kasu bakoitzerako HAD informazioa erabili gabe bi oinarri-lerro eta HAD informazioarekin hedapena eginez beste hainbat exekuzio prestatu genituen. Hain zuzen ere, ondorengo exekuzio hauekin hartu genuen parte ataza horretan (exekuzio bakoitzaren izenaren ondoren parentesi artean emaitzen taulan erabiliko ditugun identifikadoreak daude):

- Elebakarra, HADa erabili gabe:
  - `En2EnNowsd` (*exek1*): jatorrizko hitzak gaietan; jatorrizko hitzak dokumentuetan.
  - `En2EnNowsdPsrel` (*exek2*): `En2EnNowsd`ren berdina, baina PRFarekin.
- Elebakarra, HADa erabiliz:
  - `En2EnNusDocsPsrel` (*exek3*): jatorrizko hitzak gaietan; jatorrizko hitz eta hedapeneko hitzak dokumentuetan, hedapenerako NUS sistemak zehaztutako adierarik onena erabiliz; PRFarekin.
  - `En2EnUbcDocsPsrel` (*exek4*): jatorrizko hitzak gaietan; jatorrizko hitz eta hedapeneko hitzak dokumentuetan, hedapenerako UBC sistemak zehaztutako adierarik onena erabiliz; PRFarekin.
  - `En2EnFullStructTopNusDocsPsrel` (*exek5*): jatorrizko eta hedapen osoko hitzak gaietan; jatorrizko hitz eta hedapeneko hitzak dokumentuetan, hedapenerako NUS sistemak zehaztutako adierarik onena erabiliz; PRFarekin.
- Hizkuntza artekoa, HADa erabili gabe:
  - `Es2EnNowsd` (*exek6*): jatorrizko hitzak gaietan (gaztelaniazkoak); itzulitako terminoak dokumentuetan (ingelesetik gaztelaniara).
  - `Es2EnNowsdPsrel` (*exek7*): `Es2EnNowsd`ren berdina, baina PRFarekin.

exekuzioa	MAP	*exek2*rekiko ΔMAP	GMAP	*exek2*rekiko ΔGMAP
HAD gabe				
exek1	0,3534 ***	% -7,26	0,1488 ***	% -5,36
exek2	**0,3810**	———	**0,1572**	———
HADarekin				
exek3	0,3862	% 1,35	0,1541	% -2,02
exek4	**0,3899**	% 2,33	**0,1552**	% -1,29
exek5	0,3890	% 2,10	0,1532	% -2,57

4.1 taula – Ataza elebakarreko gure exekuzioen emaitza ofizialak

- Hizkuntza artekoa, HADa erabiliz:
  - `Es2EnNusDocsPsrel` (*exek8*): jatorrizko hitzak gaietan (gaztelaniazkoak); itzulitako terminoak dokumentuetan (ingelesetik gaztelaniara), itzulpenerako NUS sistemak zehaztutako adierarik onena erabiliz; PRFarekin.
  - `Es2EnUbcDocsPsrel` (*exek9*): jatorrizko hitzak gaietan (gaztelaniazkoak); itzulitako terminoak dokumentuetan (ingelesetik gaztelaniara), itzulpenerako UBC sistemak zehaztutako adierarik onena erabiliz; PRFarekin.
  - `Es2En1stTopsNusDocsPsrel` (*exek10*): itzulitako terminoak gaietan (gaztelaniatik ingelesera), lehenengo adiera erabiliz; jatorrizko hitz eta hedapeneko hitzak dokumentuetan, hedapenerako NUS sistemak zehaztutako adierarik onena erabiliz; PRFarekin.
  - `Es2En1stTopsUbcDocsPsrel` (*exek11*): itzulitako terminoak gaietan (gaztelaniatik ingelesera), lehenengo adiera erabiliz; jatorrizko hitz eta hedapeneko hitzak dokumentuetan, hedapenerako UBC sistemak zehaztutako adierarik onena erabiliz; PRFarekin.

4.5.1 Emaitza ofizialak

4.1 eta 4.2 tauletan ikus daitezke Robust-WSD atazan lortutako emaitza ofizialak, ataza elebakarrekoak eta hizkuntza arteko atazakoak, hurrenez hurren. Exekuzio bakoitzerako MAP eta GMAP balioak jarri ditugu. Gainera, ataza bakoitzerako oinarri-lerroko sistema moduan HAD gabeko exekuzio-

exekuzioa	MAP	*exek7*rekiko ΔMAP	GMAP	*exek7*rekiko ΔGMAP
HAD gabe				
exek6	0,1835 ***	% -6,22	**0,0164**	% 0,39
exek7	**0,1957**	———	0,0162	———
HADarekin				
exek8	0,2138 ***	% 9,24	0,0205 **	% 25,47
exek9	0,2100 **	% 7,32	**0,0212** ***	% 30,25
exek10	0,2350 **	% 20,06	0,0176	% 8,64
exek11	**0,2356** **	% 20,39	0,0172	% 6,17

4.2 taula – Hizkuntza arteko atazako gure exekuzioen emaitza ofizialak

rik onena (MAP altuena duena, alegia, *exek2* edo *exek7*) hartu dugu, eta horrekiko hobekuntza zenbatekoa izan den zehaztu dugu Δ zutabean. Hobekuntza horiek estatistikoki esanguratsuak diren edo ez begiratu dugu Paired Randomization Test erabiliz (ikus 3.7.2 atala), eta emaitzen tauletan * batez adierazi dugu % 90eko konfiantza-mailako esangura estatistikoa, ** % 95ekoa eta *** % 99koa. Horretaz gain, azpimultzo (HADarekin edo gabe) bakoitzeko emaitzarik onena beltzez markatuta dago.

Ikus daiteke HADa eraginkorra izan dela esperimentu hauetan. IB elebakarrari dagokionez, hedapena eginez lortzen da emaitzarik onena (*exek4* edo `En2EnUbcDocsPsrel`), baina, ez da estatistikoki esanguratsua *exek2* edo `En2EnNowsdPsrel` oinarri-lerroarekiko. Hizkuntza arteko atazan ere, hedapena eginez hobekuntzak lortu ditugu oinarri-lerroekiko. Gainera, kasu honetan, aldeak estatistikoki esanguratsuak dira. Emaitza hauek bat datoz entrenamendurako bildumarekin egindako esperimentuekin. Entrenamenduko datu-multzoaren gainean hedapen onena eginez, hedapen osoa eginez baino emaitza hobeak lortu ditugu. Eta hemen erakutsitako emaitza hauetan ez bezala, hobekuntzak estatistikoki esanguratsuak dira.

Gure helburu nagusiena ez bazen ere, gure sistema atazako sailkapen orokorrean postu onean gelditu zen. Ataza elebakarrean 8 taldek hartu zuten parte eta lehenengo 5 sistema onenen sailkapena ikus daiteke 4.3 taulan (partaide bakoitzaren exekuziorik onena kontuan hartuz). Hor ikus daitekeen moduan, HAD gabeko gure exekuziorik onena 4. postuan geratu zen, eta HADa erabilita exekuzioa 3. postuan. Hizkuntza arteko atazan, berriz, 4

postua	partaidea	esperimentua	MAP	GMAP
HAD gabe				
1.	unine	MONO-EN-TEST-CLEF2008.UNINE.UNINEROBUST4	0,4514	0,2117
2.	geneva	MONO-EN-TEST-CLEF2008.GENEVA.ISILEMTDN	0,3717	0,1653
3.	ucm	MONO-EN-TEST-CLEF2008.UCM.BM25 B01	0,3834	0,1528
4.	ixa	MONO-EN-TEST-CLEF2008.IXA.EN2ENNOWSDPSREL	0,3810	0,1572
5.	ufrgs	MONO-EN-TEST-CLEF2008.UFRGS.UFRGS R MONO2 TEST	0,3394	0,1396
HADarekin				
1.	unine	WSD-MONO-EN-TEST-CLEF2008.UNINE.UNINEROBUST6	0,4498	0,2154
2.	ucm	WSD-MONO-EN-TEST-CLEF2008.UCM.BM25 B01 CLAUSES 09	0,3957	0,1617
3.	ixa	WSD-MONO-EN-TEST-CLEF2008.IXA.EN2ENUBCDOCSPSREL	0,3899	0,1552
4.	geneva	WSD-MONO-EN-TEST-CLEF2008.GENEVA.ISINUSLWTDN	0,3813	0,1625
5.	ufrgs	WSD-MONO-EN-TEST-CLEF2008.UFRGS.UFRGS R MONO WSD5 TEST	0,3464	0,1417

4.3 taula – Ataza elebakarreko parte-hartzaile onenen emaitza ofizialak.

taldek hartu zuten parte eta horien sailkapena 4.4 taulan agertzen da. Ataza honetan 3. eta 1. postuetan geratu ginen.

Esperimentu eta emaitzen analisi bat egin ostean, dokumentuen hedapena gaien hedapena baino onuragarriagoa dela ondorioztatu dugu. Izan ere, entrenamendu-fasean kontsulten hedapena burutzeko hainbat kontsulta egituratu konplexurekin probak egin ditugun arren, dokumentuen hedapen soilarekin lortutako emaitzak hobetzea ez baitugu lortu.

Esan, esperimentu hauetan gai zailak tratatzeko ez dugula ezer berezirik egin, eta MAP hobetzea izan dela gure helburua, GMAP neurriari kasu gehiegirik egin gabe.

4.5.2 Bestelako esperimentuak

Orain arte aipatu ez ditugun beste esperimentu batzuk ere egin ditugu, hemen kontatutako esperimentuen aldaera batzuk, hain zuzen ere. Besteak beste, gaietako *title* atala bakarrik erabiliz (*desc* ez) edo lemaz gain *synset* kodeak ere erabiliz egin ditugu esperimentuak. Ea baldintza horietan hedapenak gehiago laguntzen duen ikusi nahi izan dugu esperimentu hauekin, baina, ez dugunez horrelako joera garbirik ikusi, ez ditugu emaitzak jarriko.

postua	partaidea	esperimentua	MAP	GMAP
HAD gabe				
1.	ufrgs	BILI-X2EN-TEST-CLEF2008.UFRGS.UFRGS R BI3 TEST	0,3638	0,1300
2.	geneva	BILI-X2EN-TEST-CLEF2008.GENEVA.ISIESENTD	0,3036	0,1096
3.	ixa	BILI-X2EN-TEST-CLEF2008.IXA.ES2ENNOWSDPSREL	0,1957	0,0162
4.	uniba	BILI-X2EN-TEST-CLEF2008.UNIBA.CROSS1TDNUS2F	0,0256	0,0004
5.	–	–	–	–
HADarekin				
1.	ixa	WSD-BILI-X2EN-TEST-CLEF2008.IXA.ES2EN1STTOPSUBCDOCSPSREL	0,2356	0,0171
2.	ufrgs	WSD-BILI-X2EN-TEST-CLEF2008.UFRGS.UFRGS R BI WSD1 TEST	0,2177	0,0514
3.	geneva	WSD-BILI-X2EN-TEST-CLEF2008.GENEVA.ISIESPWSDTDN	0,0970	0,0037
4.	uniba	WSD-BILI-X2EN-TEST-CLEF2008.UNIBA.CROSSWSD12NUS2F	0,0723	0,0016
5.	–	–	–	–

4.4 taula – Hizkuntza arteko atazako parte-hartzaile onenen emaitza ofizialak.

4.6 Ondorioak

Kapitulu honetan, batez ere, *Robust WSD Task @ CLEF 2008* atazan parte hartzeko egindako esperimentuak azaldu ditugu. Horretarako, HAD informazioarekin aberastutako gai- eta dokumentu-bildumez baliatu, eta kontsulta- eta dokumentu-hedapena egin ditugu. Horretarako, antolatzaileek gure esku jarri dituzten ingelesezko eta gaztelaniazko WordNetak erabili ditugu. Kanpoko baliabide hori bakarrik erabiliz hedapenak egin, eta, parametroak doitu gabe, emaitza onak eta atazako sailkapen orokorrean postu onak lortu ditugu.

Gure helburu nagusia HADa erabiliz IB atazetan emaitza hobeak lortzea zen. Eta lortu dugu, hizkuntza arteko IBan hobekuntza nabarmenak eta elebakarrean hobekuntza txikiagoak erdietsiz, nahiz eta ez izan estatistikoki esanguratsuak. Orain arteko lanik gehienak kontsulta-hedapenean zentratu badira ere, frogatu dugu dokumentu-hedapena egiteko HAD estrategia ona dela.

Honenbestez, ondorengo ikerketa-galdera hauek erantzun ahal izan ditugu:

– **IG 1** – ***Hitzen adiera-desanbiguazioan eta ezagutza-base lexikal bateko sinonimoetan oinarritutako hedapena eginez hobetzen al da IB sistemaren eraginkortasuna?***

HAD informazioarekin etiketatutako gai- eta dokumentu-bildumez baliatuz, kontsulta- eta dokumentu-hedapena egin ditugu WordNeteko sinonimoak erabiliz. Baliabide horiekin eta parametroak doitu gabe ataza elebakarrean (ingelesa) berreskurapeneko emaitzak hobetzea lortu dugu oinarri-lerroko sistemarekiko, nahiz eta lortutako hobekuntza ez izan estatistikoki esanguratsua.

- 1.1 - *Hedapen-teknika hau egokia al da kontsulta- zein dokumentu-hedapenerako? Bi hauetakoren bat ba al da bestea baino eraginkorragoa?*

Gure esperimentuetan bai kontsultak bai dokumentuak hedatu ditugu. Erabiltzaileak egin ohi dituen kontsultak nahiko motzak izan ohi dira, eta, hortaz, gerta daiteke ez edukitzea nahiko testuinguru adiera-desanbiguazioa zuzen burutzeko. Baina, erabili dugun datu-multzoko gaien *title* eta *description* eremuak erabili ditugunez, esperimentu hauetako kontsultak nahiko luzeak dira. Gainerakoan, behin HAD informazioa izanda, hedapen-prozesua berdin-berdina da kontsultentzako ala dokumentuentzako. Desberdintasuna hedapenetik lortutako hitz berri horiek IB sisteman txertatzeko moduan dago. Hainbat kontsulta egituratu konplexurekin esperimentuak egin ditugu, jatorrizko hitzak eta hedapenetik lortutakoak konbinatzeko kontsultarik eraginkorrena zein den jakiteko. Baina, emaitzarik hoberenak dokumentuen hedapena bakarrik eginez lortu ditugu.

- 1.2 - *Hedapen-teknika honen eraginkortasunean zein faktorek eduki dezakete eragina?*

Hedapen-teknika honen hainbat aldaerarekin egin ditugu probak, eta hauek dira esperimentu horietatik ateratako ondorioetako batzuk:

- Hedapen-mota: osoa (hitz bakoitzaren adiera guztien sinonimo guztietara hedatu) vs onena (hitz bakoitzaren pisurik handieneko adieraren sinonimo guztietara hedatu).
  Kontsulten hedapenean, orokorrean esanda, hedapen osoarekin emaitza hobeak lortzen dira; dokumentuen hedapenean, aldiz, hedapen onena da eraginkorrena.
- Kontsultaren luzera, kontsulta osatzeko erabiliko diren gaia-

ren eremuak: *title* vs *title+description*.
Esperimentu nagusienetan *title+description* erabili dugu, eta horrela hedapena eginez estatistikoki esanguratsuak ez diren hobekuntza txikiak lortu ditugu. *title* bakarrik erabiliz ea zer gertatzen zen ikusi nahi izan dugu. Proba horietatik ondorio garbirik ezin izan dugu atera: entrenamendurako bildumarekin ez dugu hobekuntzarik lortu, baina, testeko fasean hobekuntza handia, esanguratsua dena, lortu dugu. Hortaz, ezin esan ziurtasun osoz kontsulta motzetan (2-3 hitz) hedapenak eragin handiagoa duenik.

– Kontsulta eta indizeetan erabilitako unitatea: lema vs *synset*a.
Hedapena egitean hitz berriak sartzen dira IB prozesuan, eta, hortaz, zarata sartzeko arriskua dago, hedapen desegokia egiteagatik edo gehitutako hitz berriak polisemikoak direlako. Hori ekiditeko, hedapenean lemak erabili beharrean, *synset*ak erabiliz proba batzuk egin ditugu. *Synset*ak ez direla erabilgarriak ikusi dugu esperimentu hauetan.
– HAD sistema desberdinak: UBC vs NUS.
Bi HAD sistema desberdinen irteerekin probak egin ditugu, baina ezin izan dugu ondorio garbirik atera, esperimentu batzuetan sistema batekin lortu baititugu emaitzarik hoberenak, eta alderantziz. Hala ere, kontaketak eginez gero, NUS sistema gailentzen dela esan daiteke. Aipatu behar da, SemEval-2007ko *all-words WSD* atazan NUS sistemak UBC sistemak baino emaitza zertxobait hobeak lortu zituela (Pradhan *et al.*, 2007).

Esperimentu hauetan erabilitako IB sistemak hainbat parametro ditu, hala nola, leuntze-teknikarena, PRF teknikarenak eta hedapeneko kontsultaren pisua. Parametro hauen balio desberdinekin hainbat konbinaketa probatu baditugu ere, ez dugu balio optimoaren zantzu garbirik atera.

- 1.3 - *Hedapen-teknika hau egokia al da kontsulten eta dokumentuen itzulpena egiteko hizkuntza arteko berreskurapenean?*

WordNet hainbat hizkuntzatarako dagoenez, kontzeptu baten *synset* zenbakia zein den jakinda, oso erraz lortu daitezke kontzeptu

horri dagozkion hitzak hainbat hizkuntzatan. Hitz horiek jatorrizko hizkuntzakoak izan beharrean beste hizkuntza bateko WordNetetik hartzen baditugu, itzulpena egiten ariko gara, hedapenaz gain. Hortaz, hedapen-teknika honek kontsulta eta dokumentuak itzultzeko balio du. Itzulpen-metodo hau hizkuntza arteko ataza batean (gaztelania-ingelesa) erabili eta berreskurapeneko emaitzak hobetzea lortu dugu, estatistikoki esanguratsuak diren hobekuntzak, gainera.

5

Ahaidetasuna eta IB probabilistikoa

Kapitulu honetan ahaidetasun semantikoa erabiliz eredu probabilistikoan oinarritutako IB sistemaren eraginkortasuna hobetzeko helburuarekin egindako esperimentuak azalduko ditugu. Horretarako, ezagutza-base lexikal bat (WordNet) oinarritzat duen grafo-algoritmo baten bidez, dokumentuen hedapena egitea proposatzen dugu berauekin erlazionatutako hitzak gehituz, ondoren, hedapen hori IB sisteman txertatzeko. Esperimentuak izaera desberdineko hiru datu-multzo desberdinekin egin ditugu, eta honek hainbat analisi egiteko bidea eman digu: sendotasuna datu-multzo desberdinekiko, sendotasuna parametro-ezarpen desberdinekiko eta dokumentu luzeraren eragina, besteak beste. Gainera, teknika hau aplikatuz CLEF 2009 eta 2010 edizioetako ResPubliQA atazan hartu genuen parte, eta horietan lortutako emaitzak ere ikusiko ditugu.

5.1 Aurrekariak

Aurreko kapituluko esperimentuetatik ateratako ondorioetako bat HADean oinarritutako dokumentu-hedapena eraginkorragoa dela kontsulta-hedapena baino izan da. Hori dela eta, dokumentu-hedapena lantzearekin jarraitu nahi izan dugu.

Dokumentuen hedapena egiteko, aurreko kapituluko esperimentuetan bezala, HADa erabiltzen segi genezakeen. Baina, teknika horri muga batzuk ikusi dizkiogu: hitzez hitzeko hedapena eginez, hedapena nahiko mugatua da, hitz bakoitzaren sinonimoetara mugatu baikara. Oraingo honetan, beste

hurbilpen bat probatu nahi dugu: alde batetik, hitzez hitz egin beharrean, dokumentu osoko hitz guztiak batera kontuan hartuz egitea hedapena, eta, bestetik, sinonimiaz haratagoko beste erlazioak ere kontuan hartzea hedapena egiterakoan.

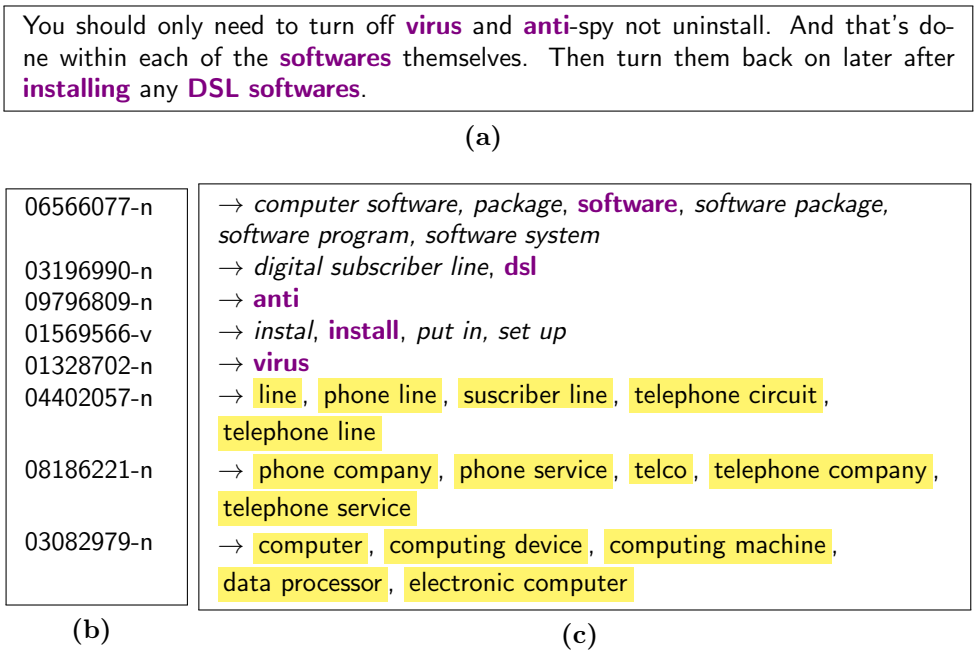
Hori honako metodo berri honen bidez egin genezakeela ikusi genuen: dokumentu osoa emanik, WordNeten oinarritutako grafo-algoritmo baten bidez dokumentuarekin ahaidetasun semantikoren bat duten kontzeptuak lortzea. Grafo-algoritmo hori erabiltzea aukeratu dugu, algoritmo hori bera hitzen antzekotasun semantikorako (Agirre *et al.*, 2009c) eta HADa egiteko (Agirre eta Soroa, 2009) erabilia izan delako, emaitza arrakastatsuak lortuz.

5.2 Ahaidetasuna dokumentuaren hedapenerako

Ahaidetasun semantikoa IBrako baliagarria den edo ez ikusteko, horretan oinarritutako dokumentu-hedapena egiteko erabil daiteke. Atal honetan hori gauzatzeko modu bat aurkeztuko dugu.

Horretarako, dokumentu bakoitzari 3.4.2 atalean azaldutako algoritmoa aplikatuko diogu WordNeten grafoaren gainean. Gogoratzeko, algoritmo honen bidez, WordNeteko kontzeptuen gainean probabilitate-banaketa gauzatzen da. Exekuzioa amaitzean, kontzeptu batek duen probabilitatea zenbat eta altuagoa izan, orduan eta ahaidetasun handiagoa izango du kontzeptu horrek landu dugun dokumentuarekin. Hortaz, probabilitate handieneko kontzeptuak hartuz, dokumentu horrekin ahaidetasun handiena duten kontzeptuak lortuko ditugu. Esperimentu gehienetan kontzeptu kopurua finkoa izango da; besterik esan ezean, ahaidetasun handieneko 100 kontzeptu hartuko ditugu hedapenerako. Hala ere, kopuru hau aldatuz IB sistemaren eraginkortasuna nola aldatzen den aztertu dugu 5.5.3 atalean.

Behin ahaidetasun handieneko kontzeptuak izanda, ingelesezko WordNet erabiliz dokumentuak hedatuko ditugu, honela: hedapenerako aukeratutako kontzeptu edo *synset* bakoitza lexikalizatzen duten *variant*ak hedapeneko hitzen multzoan gehituko ditugu. Adiera batek hizkuntza guztietako WordNetetan synset zenbaki bera du –edo mapaketa bidez erraz lor daiteke beste hizkuntza bateko *synset* zenbakia. Hortaz, hedapen hori hizkuntza batetik bestera ere gauzatu daiteke. Horretarako, *synset*ek beste hizkuntza bateko WordNetean dituzten *variant*ak hartuko ditugu. Hori izango da esperimentu



5.1 irudia – Yahoo! datu-multzoko 1005121303076 dokumentuaren hedapenaren adibidea: (a) jatorrizko dokumentua; (b) hedatuko diren kontzeptu batzuen *synset* zenbakiak; (c) hedapenetik lortutako hitzak.

batzuetan ingelesezko dokumentuak gaztelaniara itzultzeko erabiliko dugun estrategia.

5.1 irudian dokumentu baten hedapenaren adibidea ikus dezakegu. 5.1a dokumentuari grafo-algoritmoa aplikatu ondoren, besteak beste, 5.1b zerrendako *synset* zenbakia duten kontzeptuak lortuko ditugu (zerrenda hori ordenatua dago, ahaidetasun handienetik txikienera), eta horietako bakoitza lexikalizatzen duten hitzak 5.1c irudian jarri ditugu. Jatorrizko dokumentuan azaltzen diren hitzak letra lodiz eta **kolore honetan**, sinonimoak letra etzanaz eta bestelako ahaidetasunen bat duten hitzak honela markatuta jarri ditugu. Ikus daitekeen moduan, hedapenean sinonimoez gain, dokumentuan aipatzen ez diren baina erlazioren bat duten kontzeptuekin lotutako hitzak lortzen dira; adibidez, *phone company* eta *computer* hitzak.

5.3 Ahaidetasunean oinarritutako dokumentu-hedapena IB sistema baterako

Dokumentuen hedapena egin ostean, jatorri ezberdineko bi hitz-multzo lortu ditugu; alde batetik, dokumentuetan agertzen diren jatorrizko hitzak, eta, bestetik, hedapenetik lortutako hitzak. Multzo horiek mantenduz, bi indize sortu ditugu. Horrela, gure esperimentuetan egingo ditugun bilaketetan aukera izango dugu jatorrizko hitzak bakarrik erabiltzeko, edota hitz guztiak erabiltzeko, baina bi indizeen arteko bilaketa konbinatu bat eginez.

Hori dela eta, gure IB sistema inplementatzeko artearen egoerako tresna bat den, eta dokumentu-bilduma baten gainean hainbat indize sortu eta euren artean konbinatzeko aukera ematen duen testu-bilatzaile bat aukeratu dugu. Hain zuzen ere, *Managing Gigabytes for Java* (MG4J) (Boldi eta Vigna, 2005) bilatzaile librea erabili dugu atal honetan azalduko ditugun esperimentuetan.

Tresna honetan, besteak beste, BM25 ranking-funtzioa dago inplementatuta eta hori erabili dugu. 3.2.1 atalean ikusi dugu termino baten w_{Dt}^{BM25} pisua nola kalkulatzen den (3.1 ekuazioa), baita dokumentu osoaren pisua kontsultako termino guztien w_{Dt}^{BM25} pisuak batuz kalkulatzen dela ere. Hori horrela da jatorrizko hitzen indizea bakarrik erabili nahi dugunean. Jatorrizko hitzez gain hedapenetik lortutako hitzak ere kontuan hartu nahi izanez gero, bi indizeak linealki konbinatzen dira eta honela lortzen dugu Q kontsultarentzat D dokumentuak duen pisua (D' moduan adieraziko dugun dokumentuaren hedapena kontuan hartuz):

$$score(D, D', Q) = \sum_{t \in Q \cap I_D} w_{D_t}^{\text{BM25}} + \lambda \sum_{t \in Q \cap I_{D'}} w_{H_t}^{\text{BM25}} \tag{5.1}$$

non I_D eta $I_{D'}$ jatorrizko eta hedapeneko indizeak diren, t terminoa den eta λ hedapeneko indizearen pisu erlatiboa adierazteko parametro askea den.

5.4 Esperimentazio-ingurunea

Atal honetako esperimentuekin hemen aurkeztutako hedapen-teknika IBrako lagungarria eta sendoa den aztertu nahi dugu. Hori dela eta, datu-multzo desberdinen gainean egin ditugu esperimentuak, bakoitzean parametro-ezarpen desberdinak aplikatuz.

Erabilitako hiru datu-multzoak 3.5 atalean aurkeztutako Robust, ResPubliQA eta Yahoo! dira. Ikusi bezala, jatorri eta izaera desberdineko bildumak dira hirurak ere. Hau ondo datorkigu atal honetako ekarpen nagusia den *sendotasuna* frogatu ahal izateko. Sendotasuna ikuspegi desberdinetatik lor daiteke. Batetik, datu-multzoekiko sendotasuna izango dugu, zernahi datu-multzo erabilita ere, hedapenak kalterik ez badu egiten. Bestetik, analisien atalean ikusiko dugun moduan (5.5.4 atala), parametroekiko sendotasuna ere kontuan izan dugu.

3.4.2 atalean esan dugun moduan, ahaidetasun semantikorako erabili dugun algoritmoak WordNetekin lotzen ditu hedatu nahi dugun dokumentuko hitzak. Horretarako, dokumentu horretako lemak eta hauen kategoria gramatikalak behar ditu. Robust datu-multzoa adierekin etiketatuta egoteaz gain, lema eta kategoria gramatikalekin etiketatua dago. Beste bi bildumen lematizazio eta kategoria gramatikalen etiketatzea OpenNLP tresnaren bidez egin dugu[1].

3.1 eta 5.1 ekuazioetan ikusi bezala, erabilitako algoritmoan finkatu beharreko hiru parametro aske daude: k_1, b eta λ. Ondoren azalduko ditugun esperimentu bakoitzean parametro-ezarpen desberdin bat erabili dugu. Hauexek dira erabilitako parametro-ezarpenak:

- **Parametro-ezarpen lehenetsia**: Parametro guztiek beren balio lehenetsia hartuko dute. k_1 eta b parametroek 1,2 eta 0,5 balioak hartuko dituzte, hurrenez hurren, MG4J sistemako BM25 algoritmoan besterik ezeko balioak horiek dira eta. λ parametroari 0,1 jarri diogu balio lehenetsi moduan. λ parametroari 1 balioa esleituz gero, hedapenetik eratorritako hitzei jatorrizko hitzen garrantzi bera ematen ariko ginateke. Balio hori gehiegizkoa iruditu zaigunez, aipatutako balio hori lehenestea erabaki dugu.

- **Parametro-ezarpen optimizatua**: Ezarpen honetarako, datu-multzo bakoitzerako hiru parametroak optimizatu ditugu (Robertson eta Zaragoza, 2009) lanean deskribatzen duten *Promising Directions* izeneko metodoaren bidez.

5.1 taulan agertzen dira kasu bakoitzean parametro bakoitzari esleitutako balioak.

[1] http://incubator.apache.org/opennlp/

ezarpena	eredua	k_1	b	λ
lehenetsia	oinarri-lerroa	1,20	0,50	—
	hedapena	1,20	0,50	0,100
Robust	oinarri-lerroa	1,80	0,64	—
	hedapena	1,66	0,55	0,075
Yahoo!	oinarri-lerroa	0,99	0,82	—
	hedapena	0,84	0,87	0,146
ResPubliQA	oinarri-lerroa	0,09	0,56	—
	hedapena	0,13	0,65	0,090

5.1 taula – Parametro lehenetsiak eta datu-multzo bakoitzean optimizatutako parametroen balioak. λ parametroa ez da erabiltzen oinarri-lerroko sisteman.

Gure esperimentuetan hedapenak IBan duen eragina neurtzeko oinarri-lerro bat definitu dugu. Hain zuzen ere, atal honetan azaldutako esperimentuetan oinarri-lerro moduan definituko ditugu dokumentuetako jatorrizko hitzak bakarrik erabiliz egindako esperimentuak. Alegia, esperimentu horietan indize bakarra erabili dugu. Hau dela eta, 5.1 taulan ageri den moduan, oinarri-lerroan ez dago λ parametroaren beharrik (ikus 5.1 ekuazioa). Gainerako esperimentuetan hedapena aplikatu dugu eta jatorrizko hitzez gain hedapenetik lortutako hitzak erabili ditugu, bi indizeez baliatuz. Azken esperimentu hauetan lortutako emaitzak oinarri-lerroko emaitzekin alderatuz gure hedapen-metodoaren erabilgarritasuna aztertuko dugu datozen ataletan.

5.5 Emaitzak eta analisiak

3.7.2 atalean esan bezala, konparatu beharreko emaitzen arteko aldeak estatistikoki esanguratsuak diren edo ez begiratu dugu kasu bakoitzean. Datozen ataletako emaitzen tauletan * batez adieraziko dugu % 90eko konfiantza-mailako esangura estatistikoa, ** % 95ekoa eta *** % 99koa.

Tauletako lerro bakoitzeko emaitzarik onena letra lodiz markatu dugu esanguratsua denean. Eta Δ zutabean oinarri-lerroarekiko dagoen hobekuntza erlatiboa erakusten dugu.

datu-multzoa	neurria	oinarri-lerroa	hedapena	Δ
Robust	MAP	0,3781	**0,3835** ***	% 1,43
Yahoo!	MRR	0,2900	**0,2950** ***	% 1,72
	P@1	0,2142	**0,2183** ***	% 1,91
ResPubliQA	MRR	0,3931	**0,4077** ***	% 3,72
	P@1	0,2860	**0,3000** **	% 4,90

5.2 taula – Parametro lehenetsiak erabiliz lortutako emaitzak.

datu-multzoa	neurria	oinarri-lerroa	hedapena	Δ
Robust	MAP	0,3740	**0,3823** **	% 2,20
Yahoo!	MRR	0,3070	**0,3100** ***	% 0,98
	P@1	0,2293	**0,2317** *	% 1,05
ResPubliQA	MRR	0,4970	0,4942	% -0,56
	P@1	0,3980	0,3940	% -1,01

5.3 taula – Optimizatutako parametroak erabiliz lortutako emaitzak.

5.5.1 Emaitzak parametro-ezarpen desberdinekin

5.2 taulan agertzen dira parametro lehenetsiak[2] erabiliz lortu ditugun emaitzak datu-multzo bakoitzerako. Ikus dezakegun moduan, hedapenetik lortutako informazioa erabiltzean lortzen ditugun emaitzak ("hedapena" zutabea) hobeak dira oinarri-lerrokoak baino kasu guztietan, % 1,43 eta % 4,90 bitarteko hobekuntza erlatiboak lortuz.

Parametro guztiak optimizatuz egindako esperimentuen emaitzak 5.3 taulan zehaztu ditugu. Robust eta Yahoo! bildumekin egindako esperimentuetan hedapenarekin estatistikoki esanguratsuak diren hobekuntzak lortu ditugu. ResPubliQAren kasuan, berriz, ez dugu hobekuntzarik lortu, baina aldeak ere ez dira esanguratsuak.

Bi tauletako oinarri-lerroak alderatuz gero, ikusten da parametroen optimizazioak eragina duela datu-multzo guztietan. Esaterako, ResPubliQAn MRRa nabarmen hobetu da 0,3931etik 0,4970era (% 26ko hobekuntza er-

[2]Parametro lehenetsiak esatean besterik ezeko parametroak esan nahi dugu, alegia, optimizatu gabeko parametroak.

latiboa), Yahoo!-n zerbait igo da (% 5,8) eta Robusten zertxobait jaitsi (% -0,01). Alde hauen arrazoia normalean parametroen balio lehenetsiak TREC estiloko datu-multzoetan oinarrituta jartzen direla da, eta Robust izaera horretako datu-multzo bat denez, bertan optimizatutako parametro eta balio lehenetsien arteko aldea ez da hain handia (ikus 5.1 taula). ResPubliQAko dokumentuak dira TREC estiloko dokumentuekin konparatuz gero desberdinenak (askoz ere motzagoak dira), eta alde hori ere nabaritzen da parametroetan (balio lehenetsi eta optimizatutako balioen artean, ResPubliQAren kasuan dago alderik handiena). Hau dela eta, parametroak optimizatzean hobekuntzarik handiena ResPubliQAn lortu dugu. Alegia, parametroen balio lehenetsiak ez dira batere egokiak ResPubliQA datu-multzoarentzat.

5.5.2 Lambda aztertzen

Aurreko atalean ikusi dugu k_1 eta b parametroen eragina nabarmena dela IB emaitzetan. Atal honetan, hirugarren parametroaren, hau da, λren eragina zein den aztertu nahi izan dugu.

λren balio optimoek hedapenetik eratorritako terminoen erabilgarritasuna adierazten dute, eta esperimentu hauetan optimizatuta hartu dituen balioak 0,075 eta 0,146 bitartekoak izan dira (ikus 5.1 taula).

Parametro honen eragina ikusteko, beste bi parametroei (k_1 eta b) balio lehenetsi konstantea esleitu eta λ parametroari balio desberdinak emanez esperimentuak egin ditugu. Esperimentu hauen emaitzak 5.4 taulan ageri dira. Datu-multzo desberdinen emaitzak konparatuz, parametro lehenetsiak erabiltzeagatik oinarri-lerroko sisteman galera handienak izan dituen datu-multzoan (ResPubliQAn) lortzen dira hobekuntzarik handienak λ hedapenpisua bakarrik optimizatzen dugunean. Kasu horretan pisu horrek hartzen duen balioa oso altua da (0,61) beste esperimentu guztietako balioekin konparatuz (0,07-0,3 tartean dago beste kasu guztietan). Hortik ezin dugu ondorioztatu λ gero eta altuagoa bada, hobekuntza orduan eta handiagoa denik. Esan dezakegu hori goi-muga batera iritsi arte bakarrik gertatuko dela; eta kasu honetan, ResPubliQA goi-muga horretatik urrun dago (0,4221era iritsi gara, eta kasurik optimoenean 0,4942ra).

datu-multzoa	neurria	oinarri-lerroa	hedapena	Δ	λ
Robust	MAP	0,3781	**0,3881** ***	% 2,64	0,18
Yahoo!	MRR	0,2900	**0,2980** ***	% 2,76	0,27
	P@1	0,2142	**0,2212** ***	% 3,27	0,27
ResPubliQA	MRR	0,3931	**0,4221** ***	% 7,39	0,61
	P@1	0,2860	**0,3180** **	% 11,19	0,61

5.4 taula – λ parametroa bakarrik optimizatuz lortutako emaitzak. λ parametroaren balio optimo horiek zeintzuk diren ere erakusten da.

5.5.3 Hedatutako kontzeptu kopuruaren eragina aztertzen

3.4.2 atalean azaldu dugu nola lortzen dugun dokumentu batekin erlazionatutako kontzeptuak ahaidetasunaren arabera ordenatzea. Kontzeptu horietako batzuk hedatuz, dokumentuarekin erlazionatutako terminoak lortzen ditugu. Hortaz, hedapen-prozesu horretan badago beste parametro aske bat, hain zuzen ere, hedatu beharreko kontzeptu kopurua.

Parametro honen eragina ikusi nahirik, Robust datu-multzoaren gainean esperimentu batzuk egin ditugu. Esperimentu batetik bestera aldatzen den bakarra hedatutako kontzeptu kopurua da: 100, 500 edo 750. Aipatu, 100 kontzeptu hedatzean 268 hitz gehitzen ditugula hedapeneko indizean, 500 kontzepturekin 1.247 hitz, eta 750 kontzepturekin 1831 hitz. Hitz kopuruetan aldeak handiak izan arren, esperimentu hauetan lortutako emaitzen arteko aldeak ez direla esanguratsuak ikusi dugu.

Beraz, besterik esan ezean, kapitulu honetako esperimentu guztietan dokumentuarekin erlazio handiena duten 100 kontzeptu hedatu ditugu.

5.5.4 Sendotasuna aztertzen

5.5.1 atalean ikusi dugu nola parametroen optimizazioak garrantzi handia duen. Baina parametroak optimizatu ahal izateko komenigarria da entrenamendurako datuak izatea, eta datu-multzo erreal askotan ez da horrelakorik izaten. Horrelako egoera batean egongo bagina, ea beste datu-multzo batean optimizatutako parametroen balioak guk erabili nahiko genukeen datu-multzorako egokiak izan daitezkeen aztertu nahi izan dugu atal honetan azal-

datu-multzoa	parametroak	neurria	oinarri-lerroa	hedapena	Δ
Robust	lehenetsia	MAP	0,3781	**0,3835** ***	% 1,43
	Robust	MAP	0,3740	**0,3823** **	% 2,20
	Yahoo!	MAP	0,3786	0,3759	% -0,72
	ResPubliQA	MAP	0,3146	**0,3346** ***	% 6,35
Yahoo!	lehenetsia	MRR	0,2900	**0,2950** ***	% 1,72
	Robust	MRR	0,2920	0,2920	% 0,0
	Yahoo!	MRR	0,3070	**0,3100** **	% 0,98
	ResPubliQA	MRR	0,2600	**0,2750** ***	% 5,77
ResPubliQA	lehenetsia	MRR	0,3931	**0,4077** ***	% 3,72
	Robust	MRR	0,3066	**0,3655** ***	% 19,22
	Yahoo!	MRR	0,3010	**0,3459** ***	% 14,93
	ResPubliQA	MRR	0,4970	0,4942	% -0,56

5.5 taula – Beste bildumetan optimizatutako parametroak erabiliz lortutako emaitzak. Ezarpen lehenetsiko emaitzak eta datu-multzo berean optimizatutakoan lortutako emaitzak ere erakusten dira alderatzeko.

duko ditugun esperimentuekin. Horretaz gain, horrelako egoera ez-optimo batean gure sistemaren sendotasuna eta hedapen-teknikak emaitzetan zenbateko ekarpena egiten duen ere ikusi nahi izan dugu.

Horretarako, datu-multzo bakoitzaren gainean beste bildumetan optimizatutako parametroekin egin ditugu probak eta lortutako emaitzak 5.5 taulan jarri ditugu. Taula horren lehenengo zutabean IBa egiteko erabili dugun datu-multzoa agertzen da eta bigarren zutabean, berriz, erabilitako parametroak zein datu-multzoren gainean optimizatu diren adierazi da. Datu-multzo bakoitzerako parametro lehenetsiekin eta datu-multzo horretan bertan optimizatutako parametroekin egindako probak ere gehitu ditugu taula honetan emaitza guztiak modu errazean alderatzeko (azken emaitza hauek 5.2 eta 5.3 tauletan jarri ditugu lehen). Emaitzetan ikus daiteke nola dokumentuen hedapena eginez emaitzak hobeak diren beste bildumetan optimizatzen direnean parametroak, bi kasutan izan ezik: Yahoo! datu-multzoan optimizatutako parametroak erabiltzen ditugunean Robust datu-multzoarekin ari garenean, eta alderantziz. Bi kasu hauetan ordea, oinarri-lerro eta hedapenaren arteko aldeak ez dira esanguratsuak. Eta beste kasu guztietan bai.

Beraz, ikusirik kasu gehienetan gure hedapen-metodoa erabiliz emaitzetan lortzen diren hobekuntzak estatistikoki esanguratsuak direla, edota hori gertatzen ez denean aldeak ez direla esanguratsuak, kapitulu honetan aurkeztutako hedapen-teknikaren sendotasuna frogatu dugu parametro-ezarpen desberdinen gainean eta egoera ez-optimoetan.

5.5.5 Dokumentuen luzera aztertzen

5.5 taulako emaitzetan begiratzen badugu, ikusiko dugu ResPubliQA datu-multzoaren gainean lortzen direla hobekuntzarik handienak. Erabili ditugun hiru datu-multzo horietan batez besteko dokumentuen luzeran alde handiak daude (ikus 3.1 taula) eta ResPubliQA datu-multzoko dokumentuak dira motzenak. Hori dela eta, dokumentuen luzerak gure teknikaren bitartez lortzen diren hobekuntzetan eraginik ba ote duen ikusi nahiean, ondoren azalduko dugun esperimentu hau egin dugu.

Dokumenturik luzeenak dituen datu-multzoa hartu dugu, Robust datu-multzoa. Dokumentu-bilduma horretatik abiatuz, sasibilduma berri batzuk sortu ditugu dokumentu guztiak artifizialki moztuz bakoitzaren luzeraren % jakin batera, modu honetan: dokumentu bakoitzaren lehenengo x hitzak bakarrik jartzen ditugu sortzen dugun dokumentu berri horretan. Horrela, jatorrizko dokumentuen % 2,5, % 10, % 20 eta % 50eko luzera duten dokumentuz osatutako lau sasibilduma berri sortu ditugu. Ehuneko horren arabera kalkulatu dugu dokumentu bakoitzerako x hitz kopurua. Sortutako bilduma berrien batez besteko luzera berriak zeintzuk diren 5.6 taulako bigarren zutabean zehaztu dira.

Esperimentu hauetan jatorrizko Robust datu-multzoaren gainean optimizatutako parametroak erabili ditugu.

5.6 taulan agertzen diren emaitzetan ikus dezakegunez, orokorrean, zenbat eta dokumentu motzagoak izan, hedapenarekin lortzen diren hobekuntzak orduan eta handiagoak direla esan dezakegu.

5.5.6 Bestelako esperimentuak

Kapitulu honetan aztergai izan dugun dokumentuak hedatzeko teknikaren aldaera batzuk ere inplementatu ditugu. Teknika honi konplexutasun batzuk gehituz edo aldaketa txiki batzuk eginez emaitzen joera ikusi nahi izan dugunez, hasteko, Yahoo! datu-multzoarekin egin ditugu ondoren azalduko ditugun proba hauek.

jatorrizkoaren zenbatekoa	luzera (hitzak)	oinarri-lerroa	hedapena	Δ
% 2,5	13	0,0794	0,0851	% 7,18
% 10	53	0,1757	0,1833	% 4,33
% 20	107	0,2292	0,2329	% 1,61
% 50	266	0,3063	0,3098	% 1,14
% 100	531	0,3740	0,3823	% 2,22

5.6 taula – Robusteko doumentuak artifizialki moztean lortutako emaitzak (MAP). Dokumentu osoaren luzeraren % zati bat hartzen da dokumentu bakoitzerako, lehenengo zutabean adierazitakoa.

Alde batetik, indizeak sortzerakoan egin dugu aldaketa txiki bat. Dokumentuetan dauden jatorrizko hitzak eta hauetatik eratorritako hedapeneko hitzak bi multzo desberdin moduan maneiatu beharrean, denak multzo bakarrean sartuz proba egin dugu. Horrela, kasu honetan indize bakarra izango dugu bi izan beharrean. Hortaz, hitz guztiek, nola jatorrizko hitzek hala hedapenekoek, garrantzi eta pisu bera izango dute bilaketetan.

Bestetik, dokumentuaren hedapen-prozesua aldatu dugu. Orain arteko esperimentuetan dokumentuarekin erlazionatutako kontzeptuak lortu eta hauek ahaidetasunaren arabera ordenatu ondoren (antzekoena zerrendan lehena jarriaz), zerrenda horretako lehen 100 kontzeptu edo *synset* hartu eta kontzeptu bakoitzarekin lotutako termino edo *variant* guztiekin egiten dugu hedapena. Proba hauetako batean *synset* bakoitzaren *variant* guztiak hartu beharrean, horietako batzuk bakarrik aukeratuko ditugu. Izan ere, argi dago kontzeptu bat adierazterakoan termino batzuk beste batzuk baino erabilgarriagoak direla. Eta hedapena termino guztiak erabiliz egiten badugu zarata handia sartzen dugu, hain erabiliak ez diren terminoak gehitzen ari garelako. Horrela bada, *variant*ak erabileraren arabera aukeratuko ditugu, *synset* bakoitzarentzat *SemCor*ren arabera *variant* erabilienak direnak bakarrik hartuz (*SemCor*ri buruzko azalpenak 3.4.4 atalean).

Beste proba batean ere, gutxi erabiltzen diren *variant*ak baztertu nahi izan ditugu. Orain artekoetan *synset*ak ordenatu ditugu ahaidetasunaren arabera. Oraingoan *variant*ak ordenatuko ditugu modu honetan. Dokumentu batek kontzeptu (*synset*) bakoitzarekin duen ahaidetasuna neurtzen duten pisuak lortuko ditugu prozesu bera jarraituaz. Gero, *synset* bakoitzaren pisu hori bere *variant*en *SemCor*reko maiztasunarekin biderkatuko du-

gu. Eta hortik lortzen den neurri horren arabera ordenatuko ditugu *variant* guztiak. Horrela lortuko dugun zerrenda horren goiko postuetan dokumentuarekin ahaidetasun handiena duten eta erabilienak diren *variant*ak egotea lortu nahi izan dugu. Zerrenda honetako lehen 100 *variant*ak erabili ditugu hedapena egiteko.

Azken honen beste aldaera batzuk ere probatu ditugu. Esaterako, *synset* bakoitzaren ahaidetasun-pisua *variant* bakoitzaren *SemCor*reko maiztasunarekin biderkatu beharrean, *variant* bakoitzaren bildumako *idf*arekin biderkatu dugu. Eta lortutako pisu horren arabera ordenatu *variant* zerrenda, hortik lehenengo 100 terminoak hartzeko. Maiztasuna eta *idf*arekin biderkatuz ere probak egin ditugu.

Gainera, lehenengo 100 *synset* edo *variant* aukeratu beharrean, 250 aukeratuz ere egin ditugu esperimentu hauek.

Baina esperimentu hauen emaitzetan ez dugunez hobekuntza nabarmenik ikusi, ez dugu teknika gehiago lantzen jarraitu eta beste bildumekin ere ez ditugu probak egin.

5.5.7 CLEF 2009ko Robust-WSD atazako emaitzak

4. kapituluko esperimentuak CLEF 2008ko Robust-WSD atazan parte hartzeko helburuarekin egin genituela esan dugu (Agirre *et al.*, 2009a). Hurrengo urtean, 2009an, berriz ere ataza bera antolatu zen eta oraingoan ere parte hartzea erabaki genuen. Baina azken honetan ahaidetasunean oinarritutako hedapena ere, alegia, kapitulu honetan ikusitakoa, erabili nahi izan genuen esperimentuetan. Horrela bada, kontsultak hitzen adiera-desanbiguazioko informazioa erabiliz hedatu genituen, eta dokumentu-hedapena ahaidetasun semantikoan oinarrituz egin genuen. Lehiaketa honek araututa zeuzkan daten barruan ezin izan genituen esperimentuak behar bezain ondo garatu, eta, ondorioz, ataza elebakarrean (ingelesez kontsultak eta dokumentuak) hedapenarekin emaitzak hobetzerik ez genuen lortu. Hizkuntza arteko atazan ere hartu genuen parte. Ataza honetarako hainbat hurbilpen probatu genituen. Horietako batean gaztelaniazko jatorrizko kontsultak bere horretan utzi eta ingelesezko dokumentuak gaztelaniara itzuli genituen ahaidetasunean oinarritutako hedapen teknika hau erabiliz (5.2 atalean azaldu dugu nola egin daitekeen dokumentuen itzulpena). Ataza honetan itzulpena egiterakoan adierarik onena aukeratzeko HAD informazioa erabiliz emaitzak zertxobait hobetzea lortu genuen, baina aldeak ez ziren estatistikoki esanguratsuak.

Hemen esperimentu hauen emaitzak eta hauen inguruko analisiak ez di-

tugu azalduko, baina (Agirre *et al.*, 2010e) lanean aurki daitezke. Izan ere, lehiaketaren ondoren esperimentuak gehiago landu ondoren egin baititugu analisi nagusienak, eta horiek aurreko ataletan azaldu ditugu.

5.5.8 CLEF 2009 eta 2010eko ResPubliQA atazako emaitzak

5.2 eta 5.3 ataletan azaldutako oinarrizko teknika hori bera erabiliz CLEF 2009 eta CLEF 2010 ebaluazio-saioko ResPubliQA atazetan hartu genuen parte (ikus 3.5.3 atala). Jarraian saio horietan lortutako emaitzak azalduko ditugu.

ResPubliQA 2009

Kapitulu honetan azaldutako esperimentuetan erabilitako datu-multzoetako bat, ResPubliQA moduan erreferentziatzen duguna, ataza honetan parte hartzeko erabili behar zena da. Lehenik, ataza honetan hartu genuen parte. Ondoren, esperimentuak gehiago garatuz, aurreko ataletan azaldutako esperimentuak egin genituen.

Ataza honetan parte hartzerakoan, antolatzaileek zehaztutako daten barruan egin behar izan genituen esperimentuak, eta ez genuen behar adina denbora izan esperimentuak zuzen garatzeko, esaterako, parametroak optimizatu gabe egin genituen esperimentuak. Hori dela eta, ataza horretako sailkapen nagusian antolatzaileek prestatutako oinarri-lerroko sistemaren atzetik geratu ginen (Peñas *et al.*, 2009). Hala ere, bai ataza elebakarrean bai hizkuntza arteko atazan hedapena erabili genuen exekuzioetan emaitza hobeak lortu genituen hedapenik gabeko exekuzioetan baino[3]. 5.2 irudian dokumentu-hedapena eraginkorra izan zen adibide bat ikus daiteke. 5.2a galderarentzako dokumentu adierazgarrietako bat 5.2b irudikoa da, eta dokumentu horren hedapenetik lortu genituen hitzetako batzuk 5.2c irudian jarri ditugu. Ikus daitekeen moduan, hedapen horretan jatorrizko dokumentu adierazgarrian ez dauden hitz batzuk ageri dira, besteak beste, *unfavourable* eta *consequences*. Hain zuzen ere, hitz horiek galderako gako-hitzetako

[3]Bi atazatan hartu genuen parte: ingeleseko elebakarrean eta euskara-ingelesa hizkuntza artekoan. Azken honetan beste ikerketa-talde batekin elkarlanean hartu genuen parte; haiek euskarako galderak ingelesera itzultzen zituzten (Saralegi eta Lopez de Lacalle, 2010) eta guk, behin galderak ingelesez edukita, orain arte azaldutako moduan ingeleseko pasarteen berreskurapena egin genuen.

Into which plant may genes be introduced and not raise any doubts about **unfavourable consequences** for people's health?

(a) Ingelesezko 32. galdera.

Whereas the Commission, having examined each of the objections raised in the light of Directive 90/220/EEC, the information submitted in the dossier and the opinion of the Scientific Committee on Plants, has reached the conclusion that there is no reason to believe that there will be any adverse effects on human health or the environment from the introduction into maize of the gene coding for phosphinotricine-acetyl-transferase and the truncated gene coding for beta-lactamase;

(b) Dokumentu adierazgarria (jrc31998D0293-en/17).

cistron factor gene coding cryptography secret_writing acetyl acetyl_group acetyl_radical ethanoyl_group ethanoyl_radical beta_lactamase penicillinase ec eec eu europe european_community european_economic_community european_union directive directing directional guiding citizens_committee committee environment environs surround surroundings corn maize zea_mays health wellness health adverse contrary homo human human_being man adverse inauspicious untoward gamboge unfavorable **unfavourable** set_up expostulation objection remonstrance remonstration dissent protest believe light lightly belief feeling impression notion opinion reason reason_out argue jurisprudence law **consequence** effect event issue outcome result upshot

(c) Hedapenetik lortutako hitz batzuk.

5.2 irudia – ResPubliQA 2009 ataza elebakarreko galdera bat, hedapenaren bidez zuzen erantzun ahal izan genuena.

batzuk dira, eta, hori izan daiteke dokumentu-hedapena erabiliz galdera horrentzako erantzuna bilatu izanaren arrazoia, hedapenik gabe ezin izan baikenuen galdera horrentzako dokumentu adierazgarririk topatu.

Lehen esan bezala, datu-multzo honekin egindako esperimentu landuenak eta hauen emaitzak azaldu ditugu aurreko ataletan. Hori dela eta, garrantzitsuena aurreko ataletan azaldu dugunez, lehiaketa honetako emaitzak eta analisiak ez ditugu hemen aipatuko. Nahi izanez gero, (Agirre *et al.*, 2010b) lanean aurki daitezke esperimentu hauen inguruko xehetasunak.

ResPubliQA 2010

Aurreko ediziotik 2010ekora aldatu zen bakarra datu-multzoa zen. Hortaz, laburbilduz, berriz ere saio honetako ataza elebakarrean egin beharrekoa hau-

exekuzioa	k_1	b	λ
exek1 (hedapena)	0,30	0,17	0,22
exek2 (oinarri-lerroa)	0,53	0,09	—

5.7 taula – ResPubliQA-2010 atazara bidalitako 2 exekuzioetan erabilitako parametro optimizatuak. *exek1* exekuzioan hedapena erabili dugu, eta *exek2*an ez.

exekuzioa	ondo	c@1	MRR
exek1 (hedapena)	130	**0,65**	**0,6067** *
exek2 (oinarri-lerroa)	123	0,62	0,5658

5.8 taula – ResPubliQA-2010 atazako 2 exekuzioen emaitzak. *exek1* exekuzioan hedapena erabili dugu, eta *exek2*an ez.

xe zen: ingelesezko galderei erantzuna ematen zieten pasarte egokiak topatu ingelesezko dokumentuetan. Aurreko edizioko datu-multzoa prestatu genuen bezalaxe prestatu genuen datu-multzo hau ere, eta esperimentuak modu berdinean garatu genituen. Parametroak optimizatzeko aurreko edizioko testerako bilduma erabili genuen.

Hainbat IB exekuzio bidali genitzakeen saio honetara antolatzaileek eskuz ebalua zitzaten. Guk ataza elebakarreko bi exekuzio bidali genituen. Bi exekuzio horien arteko desberdintasun bakarra da lehenengoan hedapena aplikatzen genuela eta bigarrenean ez. 5.7 taulan ikusten da hedapena aplikatu genuen exekuzio horretarako λ parametroak hartu zuen balio optimizatua (0,22), eta baita beste bi parametroen balio optimoak ere bi exekuzioetarako.

Ataza honetan 200 galdera erantzun behar ziren, eta, 5.8 taulan ikus daitekeen moduan, horietatik 130 ondo erantzun genituen lehenengo exekuzioan, eta 123 bigarrenean. Alegia, dokumentuen hedapena aplikatutako exekuzioan emaitza hobeak lortu genituen. Horixe bera erakusten dute taulan ageri diren c@1 eta MRR neurriek ere. Galdera-multzo horretatik 121 galderei bi exekuzioek eman diete erantzun ona. Bigarren exekuzioan gaizki erantzun, baina lehenengoan asmatu diren galderak 9 dira. Eta, alderantziz, lehenak asmatu ez, baina bigarrenak ondo erantzundako galderak, berriz, 2 dira. Esperimentu hauen xehetasun gehiago (Agirre *et al.*, 2010a) lanean aurki daitezke. Hor bertan azaltzen ditugu hizkuntza arteko atazarako egindako

esperimentuak ere.

Hortaz, datu-multzo honetan ere, dokumentuen hedapena IBrako baliagarria dela ikusi genuen. Gainera, edizio honetan sailkapen orokorrean postu hobea lortu genuen, parte-hartzaile guztion artean 16 exekuzio bidali genituen ingeleseko ataza elebakarrera, eta gure exekuziorik onena 4. postuan gelditu zen, antolatzaileek prestatutako oinarri-lerroaren pareko (Peñas *et al.*, 2010).

5.6 Ondorioak

Kapitulu honetan aurkeztu dugun teknika berritzailea da IBrako dokumentuaren hedapenari dagokionez. Teknika WordNeten oinarritutako grafo-algoritmo baten bidez dokumentuarekin erlazionatutako kontzeptuak lortzean datza. Behin kontzeptuak lortuta, hauei dagozkien terminoekin dokumentuak hedatzen ditugu. Horrela, bi indize sortzen ditugu IBko bilaketetan erabili ahal izateko; bata dokumentuko jatorrizko terminoekin, eta bestea, hedapenetik eratorritako terminoekin.

Dokumentuen hedapena egiteko prozesua nahiko pisutsua bada ere, indizeak sortzeko garaian egiten denez, behin bakarrik egiten da eta ez du denbora galtzerik ekartzen galderak egikaritzeko garaian. Gainera, dokumentu batean testuingurua galderetan izaten dena baino zabalagoa da normalean, eta, beraz, hedapena egiteko informazio gehiago dugu galderen hedapenean baino.

Ikusitakoaren arabera, hasieran planteatutako ikerketa-galderei honela erantzun diegu:

– **IG 2** – ***Ezagutza-base lexikal bidezko ahaidetasun semantikoan oinarritutako hedapena eginez hobetzen al da IB sistemaren eraginkortasuna?***

WordNeten oinarritutako grafo-algoritmo baten bidez dokumentu bakoitzarekin semantikoki erlazionatutako kontzeptuak lortu eta ahaidetasun handieneko kontzeptuak lexikalizatzen dituzten hitzekin dokumentuak hedatu ditugu. Dokumentu hedatu horiek erabiliz, bi indize sortu ditugu eredu probabilistiko klasikoan oinarritutako IB sistema bateko bilaketetan erabili ahal izateko. Hiru datu-multzo desberdinetan parametro-ezarpen desberdinak aplikatuz (parametro lehenetsiak, parametro optimizatuak eta beste

datu-multzo batean optimizatutako parametroak), hedapena erabiltzean estatistikoki esanguratsuak diren hobekuntzak lortu ditugu kasurik gehienetan.

- 2.2 - *Hedapen-teknika hau egokia al da kontsulta- zein dokumentu-hedapenerako?*

Kapitulu honetako esperimentuetan hedapen-teknika hau dokumentuen hedapenerako erabili dugu, eta, esan bezala, hobekuntzak lortu ditugu hedapena aplikatutako IB esperimentuetan. Hortaz, dokumentuen hedapenerako behintzat, egokia dela esan dezakegu.

- 2.4 - *Hedapen-teknika honen eraginkortasunean zein faktorek eduki dezakete eragina?*

Hedapen-teknika honen sendotasuna eta honen eraginkortasunean eragin dezaketen faktoreak aztertu nahi izan ditugu. Jarraian aztertutako faktore desberdinak eta proba horietatik ateratako ondorioak zerrendatuko ditugu:

– Parametroak orokorrean.
Guk proposatutako IB ereduan hainbat parametro ditugu, berezko IB ereduko k_1 eta b, eta guk gehitutako λ hedapen-pisua. Parametroek eragin handia dute guk proposatutako hedapen-eredu horren gainean, oinarri-lerroko siteman duten bezalaxe. Esan daiteke, orokorrean, parametroak optimizatutakoan lortzen direla emaitzarik onenak. Aldiz, oro har, guk proposatutako hedapen-sistema eraginkorragoa da —beti ere oinarri-lerroko sistemarekin alderatuz— parametroak optimoak ez diren kasuetan. Honen atzean dagoen azalpena honakoa da: parametroak optimoak ez direneko kasuetan emaitzetan sortzen diren galerak, hedapenarekin konpentsatzen saiatzen da sistema. Horrela, Robusten kasuan, adibidez, parametro lehenetsiekin hedapena eginez lortzen da emaitzarik onena. Antzeko zerbait ikusten da λ bakarrik optimizatuz egindako esperimentuetan: datu-multzo desberdinak konparatuz, parametro lehenetsiak erabiltzeagatik oinarri-lerroko sisteman galera handienak izan dituen datu-multzoan (ResPubliQAn) lortzen dira hobekuntzarik handienak λ hedapen-pisua optimiza-

tuz. Kasu horretan pisu horrek hartzen duen balioa oso altua da (0,61) beste esperimentu guztietako balioekin konparatuz (0,07-0,3 tartean dago beste kasu guztietan). Ezaugarri hau —alegia, parametroak optimoak ez diren kasuetan eraginkorra izatea— interesgarria da oso. Izan ere, datu-multzo berri batekin lan egin behar dugunean ez dugu izaten parametroak optimizatu ahal izateko behar adina datu askotan.

– Hedatutako kontzeptu kopurua.
  Dokumentuen hedapena egiteko kontzeptu kopuru desberdinak erabiliz proba batzuk egin ditugu. Aukeratutako kontzeptu kopuru horien arteko aldeak eta hauetatik eratorritako hitz kopuruen arteko aldeak handiak badira ere —100, 500 eta 750 kontzeptu hartuz, 268, 1.247 eta 1.831 hitz, hurrenez hurren—, esperimentu hauetan lortutako emaitzen arteko aldeak ez dira esanguratsuak.
– Dokumentuen luzera.
  Batez besteko dokumentu-luzera desberdinetako sasibilduma batzuekin esperimentuak eginez, zenbat eta dokumentu motzagoak izan, hedapen-teknika orduan eta eraginkorragoa dela ikusi dugu, salbuespenak salbuespen.

Hauez gain, 5.5.6 atalean azaldu ditugu hedapen-teknika honen beste aldaera batzuekin egindako probak. Baina, emaitzetan hobekuntza nabarmenik ez dugunez ikusi, hemen ez ditugu aipatu.

- 2.5 - *Hedapen-teknika hau egokia al da kontsulten eta dokumentuen itzulpena egiteko hizkuntza arteko berreskurapenean?*

WordNet hainbat hizkuntzatarako dagoenez, kontzeptu baten *synset* zenbakia zein den jakinda, oso erraz lor daitezke kontzeptu horri dagozkion hitzak hainbat hizkuntzatan. Hitz horiek jatorrizko hizkuntzakoak izan beharrean beste hizkuntza bateko WordNetetik hartzen baditugu, itzulpena egiten ariko gara, hedapenaz gain. Hortaz, hedapen-teknika honek kontsulta eta dokumentuak itzultzeko balio du. Itzulpen-metodo hau erabili dugu Robust-WSD 2009 atazako hizkuntza arteko atazarako (gaztelania-ingelesa) prestatutako esperimentuetan.

Oro har esanda, baliabide lexikal bat erabiliz dokumentuen hedapena egin eta IBko emaitzak hobetzea lortu dugu. Gainera, proposatutako siste-

mak sendotasuna baduela esan dezakegu. Hala ere, hainbat ikerketa-lerro irekita geratzen dira. Esaterako, guk WordNet erabili dugu baliabide lexikal moduan. Baina beste baliabideren bat, Wikipedia esaterako, edo domeinu zehatz bateko ontologiaren bat txertatzeko aukera ematen du erabili dugun ahaidetasuna neurtzeko grafo algoritmoak. Gainera, proposatu dugun dokumentu-hedapenean prozesu sinple bat jarraitzen dugunez, hedapenetik lortutako informazio guztia IB sisteman modu landuago batean txertatuz emaitza hobeak lortzen ote diren aztertu nahiko genuke. Aukera bat BM25F berreskurapen-eredu probabilistikoarekin (Robertson *et al.*, 2004) esperimentatzea da.

6

Ahaidetasuna eta hizkuntza-ereduetan oinarritutako IBa

Kapitulu honetan ahaidetasun semantikoa erabiliz hizkuntza-ereduetan oinarritutako IB sistemaren eraginkortasuna hobetzeko helburuarekin egindako esperimentuak azalduko ditugu. Horretarako, aurreko kapituluan proposatutako teknika berarekin (WordNet oinarritzat duen grafo-algoritmoa), kontsulta eta dokumentuen hedapena egitea proposatzen dugu berauekin erlazionatutako hitzak gehituz, ondoren, hedapen horiek IB sisteman txertatzeko. Oraingo honetan ere izaera desberdineko hiru datu-multzo desberdinekin egin ditugu esperimentuak. Datu-multzo desberdinen arteko analisiak eta beste hainbat analisi egiteaz gain, gure hedapen-ereduekin lortzen ditugun emaitzak artearen egoerako kontsulta-hedapen ezagun batekin, hain zuzen ere, *pseudo relevance-feedback*arekin, lortzen diren emaitzekin konparatu ditugu.

6.1 Aurrekariak

Aurreko kapituluko esperimentuetan ahaidetasun semantikoan oinarritutako hedapena eginez emaitza onak lortu ditugu. Baina bakarrik dokumentu-hedapena eginez probatu dugu, eta orain, kontsulten hedapenerako ere hain baliagarria den edo ez ikusi nahi dugu.

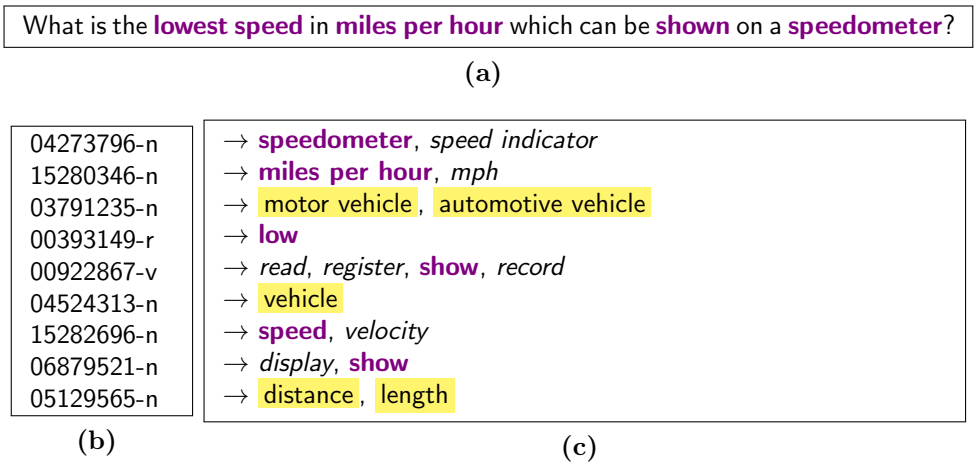
Aurreko kapituluko esperimentuetan erabili dugun MG4J sistemak kontsultetan eragile boolearrak erabiltzeko aukera ematen badu ere, kontsulta egituratu konplexuagoak egiteko edo terminoei pisu desberdinak esleitzeko aukerarik ez du eskaintzen. Kontsulta egituratuen bidez kontsultako termi-

noen artean ezar daitezkeen erlazioek eta pisuek IB sistema baten eraginkortasunean duten eragina aztertu zuten Kekäläinen eta K. Järvelin-ek (1998), eta kontsulta egituratu konplexuak eraginkorrak direla ondorioztatu zuten. Gure kontsulten hedapenetik hitz asko lortuko ditugunez, kontsulta egituratuak erabiltzea komenigarria izan daitekeela uste dugunez, MG4J beharrean, kontsulta egituratu konplexuak egiteko aukera ematen digun beste IB sistema bat erabiltzea pentsatu dugu. Eta, berriz ere, 4. kapituluko esperimentuetan erabilitako Indri sistema erabili dugu.

6.2 Ahaidetasuna testuaren hedapenerako

5.2 atalean azaldu dugu ahaidetasun semantikoaren bidez nola gauzatzen dugun dokumentu-hedapena. Hurrengo ataletako esperimentuetan ere modu berean hedatuko ditugu dokumentuak. Eta teknika hori bera erabiliko dugu kontsultak hedatzeko ere. Lehen aipatutako atal horretan azaldu dugun metodoa hauxe da, laburturik: 3.4.2 atalean aurkeztutako grafo-algoritmoa eta WordNet erabiliz, kontsulta edo dokumentu bakoitzarekin ahaidetasun handiena duten WordNeteko kontzeptuak lortuko ditugu. Behin kontzeptuak izanda, horiek lexikalizatzen dituzten hitzekin hedatuko dugu kontsulta edo dokumentua. Besterik esan ezean, ahaidetasun handieneko 100 kontzeptu hartuko ditugu hedapenerako.

Metodo honen bidez gauzatutako dokumentu-hedapen baten adibidea aurreko kapituluko 5.1 irudian ikus daiteke. Laburki, *virus*, *software* eta *DSL* aipatzen dituen dokumentu baten hedapenetik, besteak beste, *digital subscriber line*, *phone company* eta *computer* hitzak lortzen dira. 6.1 irudian, berriz, kontsulta baten hedapenaren adibidea ikus dezakegu. 6.1a kontsultari grafo-algoritmoa aplikatu ondoren, besteak beste, 6.1b zerrendako *synset* zenbakia duten kontzeptuak lortuko ditugu (zerrenda hori ordenatua dago, ahaidetasun handienetik txikienera), eta horietako bakoitza lexikalizatzen duten hitzak 6.1c irudian jarri ditugu. Jatorrizko kontsultan azaltzen diren hitzak letra lodiz eta **kolore honetan**, sinonimoak letra etzanaz eta bestelako ahaidetasunen bat duten hitzak honela markatuta jarri ditugu. Horrela bada, hedapenean kontsultan agertzen ez diren baina zerikusia duten *vehicle* eta *distance* hitzak agertzen dira.



6.1 irudia – ResPubliQA datu-multzoko 91. galderaren hedapenaren adibidea: (a) jatorrizko kontsulta; (b) hedatuko diren kontzeptu batzuen *synset* zenbakiak; (c) hedapenetik lortutako hitzak.

6.3 Ahaidetasunean oinarritutako hedapen-ereduak IB sistema baterako

Atal honetan ahaidetasunean oinarrituz garatutako dokumentuen eta kontsulten hedapen-ereduak aurkeztuko ditugu.

6.3.1 Ahaidetasunean oinarritutako dokumentu-hedapena IBrako

Ahaidetasunean oinarritutako dokumentuen hedapen-eredua gauzatzeko lehen urratsa dokumentu bakoitzaren ahaidetasun handieneko hitzak lortzea da. Hori aurreko atalean azaldutako moduan egingo dugu. Ondoren, jatorrizko hitzak eta hedapenetik eratorritako hitz horiek indize desberdinetan sartuko dira.

Hemen proposatuko dugun ahaidetasunean oinarritutako dokumentu-hedapenaren eredu honetan (*Relatedness based Document Expansion*etik RDE moduan laburtuko dugu) hainbat hizkuntza-ereduren konbinazioaren estimazioa egingo dugu, dokumentu bateko jatorrizko hitz eta dokumentu horren hedapenetik sortutako hitzetatik eratorritako hizkuntza-ereduetan oinarrituta. Dokumentu batetik kontsulta sortzeko probabilitatearen arabera sail-

katuko dira dokumentuak (Ponte eta Croft, 1998), eta probabilitate hori dokumentuaren bi errepresentazioen *query likelihood* probabilitatearen konbinazioa izango da:

$$P_{RDE}(Q \mid \Theta_{RDE}) = P(Q \mid \Theta_D)^w P(Q \mid \Theta_{D'})^{1-w} \tag{6.1}$$

non Θ_D eta $\Theta_{D'}$ jatorrizko dokumentu eta hedatutako dokumentuen errepresentazioetatik estimatutako hizkuntza-ereduak diren, hurrenez hurren, eta w jatorrizko dokumentuaren hizkuntza-ereduari esleitutako pisua den, [0..1] tartean finkatu beharrekoa.

*Query likelihood*aren estimazioa banaketa multinomialari jarraituz honela kalkulatuko dugu (jatorrizko dokumentuaren eredua erakutsiko badugu ere, hedapenarena modu berean kalkulatzen da):

$$P(Q \mid \Theta_D) = \prod_{i=1}^{|Q|} P(q_i \mid \Theta_D)^{\frac{1}{|Q|}} \tag{6.2}$$

non Q kontsultako terminoa den q_i eta $|Q|$ kontsulta horren luzera den (termino kopurua). Eta Dirichlet leuntze-teknika (Zhai eta Lafferty, 2001a) jarraituz honako hau dugu:

$$P(q_i \mid \Theta_D) = \frac{tf_{q_i D} + \mu \frac{tf_{q_i C}}{|C|}}{|D| + \mu} \tag{6.3}$$

non $tf_{q_i D}$ eta $tf_{q_i C}$ D dokumentuko eta bilduma osoko q_i kontsulta-terminoaren maiztasunak diren, hurrenez hurren, eta μ leuntze-teknikaren parametro askea den.

6.3.2 Ahaidetasunean oinarritutako kontsulta-hedapena IBrako

Ahaidetasunean oinarritutako kontsulta-hedapeneko eredu honetan (RQE moduan izendatuko dugu, *Relatedness based Query Expansion*en laburdura moduan), hasteko, 6.2 atalean azaldutako hedapen-teknikaren bitartez lortutako terminoekin hedatuko dugu kontsulta bakoitza. Honela, hedatutako kontsulta horretan kontsultako jatorrizko hitzez gain, hedapeneko terminoak izango ditugu, eta hedatutako kontsulta horretan oinarrituta berreskuratuko

ditugu dokumentuak. Dokumentu batetik kontsulta hedatu osoa (Q_{RQE}) sortzeko probabilitatearen arabera sailkatuko dira dokumentuak, probabilitate hori honela kalkulatuko delarik:

$$P_{RQE}(Q_{RQE} \mid \Theta_D) = P(Q \mid \Theta_D)^w P(Q' \mid \Theta_D)^{1-w} \tag{6.4}$$

non w jatorrizko kontsultari esleitutako pisua den ([0..1] tartean finkatu beharrekoa) eta Q kontsultaren hedapena Q' den. $P(Q \mid \Theta_D)$ *query likelihood* probabilitatea, berriz ere, banaketa multinomiala eta Dirichlet leuntzeteknika jarraituz kalkulatuko da, (6.2) eta (6.3) ekuazioetan zehaztu bezala. Hedapen-terminoak sortzeko probabilitatea, berriz, honela kalkulatzen da:

$$P(Q' \mid \Theta_D) = \prod_{q'_i}^{|Q'|} P(q'_i \mid \Theta_D)^{\frac{w_i}{W}} \tag{6.5}$$

non q'_i hedapen-terminoa den, $W = \sum_{i=1}^{|Q'|} w_i$ eta w_i hedapen-terminoari eman diogun pisua den. Pisu hau Q jatorrizko kontsultaren eta hedapen-terminoaren arteko ahaidetasun bezala ikus dezakegu eta honela kalkulatuko dugu:

$$w_i = P(q' \mid Q) = \sum_{i=1}^{N} P(q' \mid c_i) P(c_i \mid Q) \tag{6.6}$$

non c hedapen-algoritmoak itzulitako kontzeptua den, N hedapenerako hartzen ditugun kontzeptu kopurua den, $P(q' \mid c_i)$ *SemCor*eko adiera-maiztasunak erabiliz zenbatetsiko den (hots, kontsultako terminoa den q' zenbat aldiz agertzen den c_i adierarekin), eta $P(c_i \mid Q)$ aipatutako hedapen-algoritmo horrek c_i kontzeptuari emandako ahaidetasun-pisua den.

6.4 Esperimentazio-ingurunea

Aurreko atalean aurkeztutako hedapen-eredu horiek hiru datu-multzo desberdinen gainean ebaluatu ditugu, 3.5 atalean deskribatutako datu-multzoak, hain zuzen ere.

3.4.2 atalean esan dugun moduan, ahaidetasun semantikorako erabili dugun algoritmoak WordNetekin lotzen ditu hedatu nahi dugun dokumentuko hitzak. Horretarako, dokumentu horretako lemak eta hauen kategoria gramatikalak behar ditu. Robust datu-multzoa adierekin etiketatuta egoteaz

gain, lema eta kategoria gramatikalekin etiketatua dago. Beste bi bildumen lematizazio eta kategoria gramatikalen etiketatzea OpenNLP tresnaren bidez egin dugu[1].

Kapitulu honetako esperimentuak garatzeko Indri bilatzailea erabili dugu (Strohman *et al.*, 2005) (begiratu 3.3.1 atalean tresna honen inguruko zehaztapenak).

Kapitulu honetan proposatu ditugun hedapen-ereduak informazio-berreskurapenerako erabilgarriak diren edo ez aztertu nahi genuen. Horretarako, hainbat esperimentu egin ditugu eredu horiek artearen egoerako beste eredu batzuekin konparatu ahal izateko. Bi oinarri-lerro hartu ditugu. Horietako bat Indri sisteman eredu lehenetsia den *query likelihood* (QL) hizkuntza-eredua da. Oinarri moduan hartu dugun beste eredua *pseudo-relevance feedback* (PRF) deitzen dena da. Kasu honetan ere, Indri sisteman inplementatuta dagoena erabili dugu. Hain zuzen ere, Lavrenko-k proposatutako adierazgarritasun-ereduaren (ingelesezko *relevance model*) aldaera bat da (Lavrenko eta Croft, 2001), non berreskurapena egiteko erabiltzen den kontsulta jatorrizkoaren eta hedatutako kontsultaren konbinazio bat den, (6.4) ekuazioan ikusi dugunaren analogoa. Bi eredu hauetarako ere Dirichlet leuntze-teknika aukeratu dugu. Uste dugu QL eta PRF eredu sendoak izanik, oinarri-lerro onargarriak direla.

Orain arte aipatutako eredu hauek guztiek hainbat parametro aske dituzte. PRF ereduak 3 parametro ditu: dokumentu (d) eta termino (t) kopurua, eta w pisua (cf. (6.4) ekuazioa). RDE ereduak ere w parametroa du (cf. (6.1) ekuazioa). RQE ereduak 2 parametro ditu: w pisua (cf. (6.4) ekuazioa) eta N, hedapenean erabiliko den kontzeptu kopurua zehazten duena ((6.6) ekuazioa). Hauetaz gain, eredu guztietarako Dirichlet leuntze-teknika erabili dugu, eta honek μ leuntze-parametroa du. Datu-multzo bakoitzeko entrenamendurako bilduma erabiltzen dugu parametro hauen balioak *grid* algoritmo sinple baten bidez doitu eta finkatzeko. μ parametroarentzat [100,1200] tarteko 100nakako balioak probatu genituen ResPubliQA eta Yahoo! datu-multzoentzat, eta [100,2000] tartekoak Robust datu-multzoarentzat. w parametroarentzat [0,1] tarteko balioekin probatu genuen. d parametroa [2,50] eta t eta N [1,200] tarteko balioekin probatu genituen (tarte horietako 10 balio ezberdinekin probatu genituen). Doiketa horietan eredu eta datu-multzo bakoitzerako MAP altuena lortzen duten parametroen balioak, hots, gure esperimentuetan erabili ditugun balioak 6.1 taulan ikus daitezke.

[1] http://incubator.apache.org/opennlp/

	QL	PRF				RDE		RQE		
datu-multzoa	μ	μ	d	t	w	μ	w	μ	N	w
Robust	1000	1000	10	50	0,3	1200	0,8	2000	100	0,5
Yahoo!	200	200	2	20	0,8	200	0,8	200	50	0,7
ResPubliQA	100	100	10	30	0,8	100	0,7	100	125	0,7

6.1 taula – Erabilitako parametro askeen balioak datu-multzo bakoitzerako.

6.5 Emaitzak eta analisiak

Atal honetan QL eta PRF oinarri-lerroekin eta guk proposatutako RDE eta RQE ereduekin lortutako emaitzak aurkeztuko ditugu.

Gogoratu, 3.7.2 atalean esan bezala, konparatu beharreko emaitzen arteko aldeak estatistikoki esanguratsuak diren edo ez begiratu dugula Paired Randomization Test erabiliz eta datozen ataletako emaitzen tauletan * batez adierazi dugula % 90eko konfiantza-mailako esangura estatistikoa, ** ikurrez % 95ekoa eta *** ikurrez % 99koa.

6.5.1 Emaitza nagusiak

QLarekin alderatuz

Gure esperimentuetako emaitza nagusiak 6.2 taulan ageri dira. Taula horretako emaitzen lehenengo zutabean QLarekin lortutako emaitzak ikus ditzakegu eta hurrengo zutabean PRFarekin lortutakoak. Horren ondoren bi eredu hauen arteko aldeak jarri ditugu. PRFko emaitzak nahasiak dira. Izan ere, Robust datu-multzoarentzat oso eraginkorra da, hobekuntza handiak lortu baititugu, batez ere, MAP neurrirako. Eta P@5 neurrian izan ezik, hobekuntza guztiak estatistikoki esanguratsuak dira. Baina, aldiz, beste bi bildumetan PRFarekin ez dugu horrelako hobekuntzarik lortu. Yahoo! datu-multzoan hobekuntza txiki batzuk lortu ditugu MRR eta P@10 neurrietan (ez dira estatistikoki esanguratsuak), eta P@5 neurrian emaitzak okerragoak dira. ResPubliQA datu-multzoan PRFarekin emaitzak txarragoak dira, batez ere, MRR neurriarentzat.

6.2 taulako azken bi zutabe-multzotan RDE eta RQE ereduekin lortutako emaitzak ikus ditzakegu. QL ereduarekiko lortutako aldeak ere jarri ditugu,

		QL	PRF		RDE		RQE	
datu-mult.	neurria	emaitza	emaitza	Δ QL	emaitza	Δ QL	emaitza	Δ QL
Robust	MAP	0,3322	**0,3669** ***	% 10,44	**0,3387** **	% 1,95	**0,3367**	% 1,36
	GMAP	0,1321	**0,1438** ***	% 8,90	**0,1351**	% 2,26	**0,1434** **	% 8,59
	P@5	0,4250	**0,4363**	% 2,65	**0,4300**	% 1,18	0,4225	% -0,59
	P@10	0,3531	**0,3738** ***	% 5,84	**0,3556**	% 0,71	**0,3581**	% 1,42
Yahoo!	MRR	0,2636	**0,2640**	% 0,15	**0,2752** ***	% 4,42	**0,2722** ***	% 3,26
	P@5	0,0667	0,0663 **	% -0,56	**0,0691** ***	% 3,64	**0,0688** ***	% 3,21
	P@10	0,0395	**0,0396**	% 0,25	**0,0412** ***	% 4,29	**0,0410** ***	% 3,91
ResPubl.	MRR	0,4877	0,4633 ***	% -5,00	**0,4926**	% 1,02	**0,4978**	% 2,07
	P@5	0,1244	0,1200 *	% -3,54	0,1236	% -0,64	**0,1268**	% 1,93
	P@10	0,0680	0,0678	% -0,29	**0,0694**	% 2,06	0,0678	% -0,29

6.2 taula – Eredu guztietako emaitzak. Δ zutabeetan QLarekiko hobekuntza erlatiboak agertzen dira. QL baino hobeak diren emaitzak lodiz markatuak daude.

eta ikus daiteke nola bi ereduek QL hobetzen duten ia datu-multzo eta neurri guztietarako. Yahoo! datu-multzoarentzat hobekuntza nabarmenak lortu ditugu.

PRFarekin alderatuz

6.3 taulan, berriz ere, PRF, RDE eta RQE ereduekin lortutako emaitzak errepikatzen dira, PRFaren emaitzekin modu erraz batean alderatu ahal izateko.

Emaitzarik onenak lotzen dituen eredua datu-multzo bakoitzerako desberdina dela ikus dezakegu: Robusterako PRFa, Yahoo!rako RDEa eta ResPubliQArako RQEa. Gehienetan aldeak estatistikoki esanguratsuak dira.

Jakina da PRFa ondo dabilela dokumentu-bilduma eta galdera jakin batzuetarako, baina ez horren ondo beste batzuetarako (Ruthven eta Lalmas, 2003). Emaitzen tauletan GMAP neurria ere zehaztu dugu Robust datu-multzorako (beste datu-multzoetan ez da adierazgarria neurri hau). GMAP neurriak kontsulta guztietan ondo egiten duten sistemen alde egiten du, kontsulta batzuetarako oso ondo baina beste batzuetarako hain ondo ere ez dabiltzan sistemak kaltetuaz (Robertson, 2006). Kontuan hartzekoa da, esaterako, RQEak PRFak baino MAP baxuagoa lortzen duela, baina ia GMAP berdina lortzen duela. Ildo honetatik jarraituz, hurrengo atalean kontsulta bakoitzaren emaitzen analisia egingo dugu eta irudietan ikusiko dugu RDE eta RQE ereduak PRFra nahikoa hurbiltzen direla, eta hauen emaitzak hobeak direla

		PRF	RDE		RQE	
datu-multzoa	neurria	emaitza	emaitza	Δ PRF	emaitza	Δ PRF
Robust	MAP	0,3669	0,3387 ***	% -7,69	0,3367 ***	% -8,22
	GMAP	0,1438	0,1351 **	% -6,10	0,1434	% -0,29
	P@5	0,4363	0,4300	% -1,43	0,4225	% -3,15
	P@10	0,3738	0,3556 ***	% -4,85	0,3581 *	% -4,18
Yahoo!	MRR	0,2640	**0,2752** ***	% 4,26	**0,2722** ***	% 3,11
	P@5	0,0663	**0,0691** ***	% 4,22	**0,0688** ***	% 3,79
	P@10	0,0396	**0,0412** ***	% 4,03	**0,0410** ***	% 3,65
ResPubliQA	MRR	0,4633	**0,4926** ***	% 6,33	**0,4978** ***	% 7,44
	P@5	0,1200	**0,1236**	% 3,00	**0,1268** ***	% 5,67
	P@10	0,0678	**0,0694**	% 2,36	0,0678	% 0,00

6.3 taula – PRF, RDE eta RQE ereduen emaitzak. Δ zutabeetan PRFarekiko hobekuntza erlatiboak agertzen dira. PRF baino hobeak diren emaitzak lodiz markatuak daude.

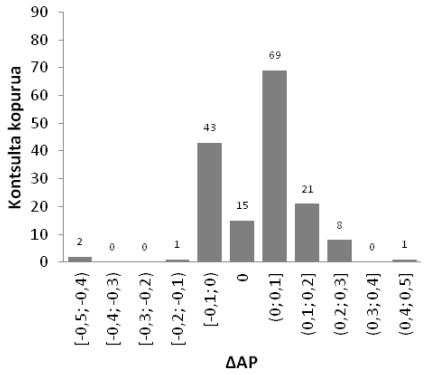
kontsulta jakin batzuetan, kontsulta zailenetan zehatzago esateko.

6.5.2 Kontsulten banakako analisiak

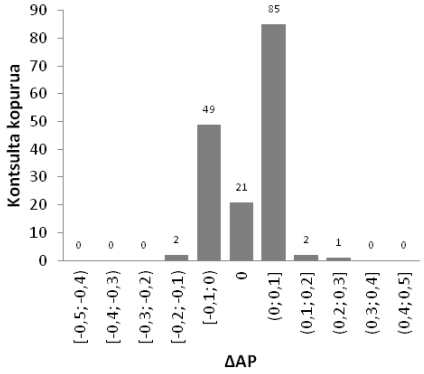
Aurreko atalean aurkeztutako emaitzetan ikusten da datu-multzo bakoitzerako eredu batetik bestera alde nabarmenak daudela. Atal honetan, alde horiek aztertzen jarraituko dugu, baina emaitza orokorrak konparatu beharrean, kontsulta bakoitzaren emaitzak banaka konparatuko ditugu.

Alde batetik, bi sistema hartu eta kontsulta bakoitzarentzat sistema batetik besterako AP (*average precision*) diferentziak kalkulatu, diferentzia horien arabera ordenatu eta multzokatu eta multzo horiek irudikatu ditugu 6.2, 6.3 eta 6.4 irudietan, multzo bakoitzean zenbat kontsulta dauden adieraziz (X ardatzean diferentzia-tarteak jarri ditugu eta Y ardatzean tarte bakoitzeko kontsulta kopurua). Diferentzia horiek QL eta PRF oinarri-lerroekiko kalkulatu dira. Horrela, irudi horietan QL edo PRF ereduekiko beste ereduek lortutako AP diferentziak ikus ditzakegu. Diferentzia hori positiboa bada, kontsulta hori kasuan kasuko oinarri-lerro horrek baino hobeto erantzun duela adierazi nahi du, eta alderantziz negatiboa bada.

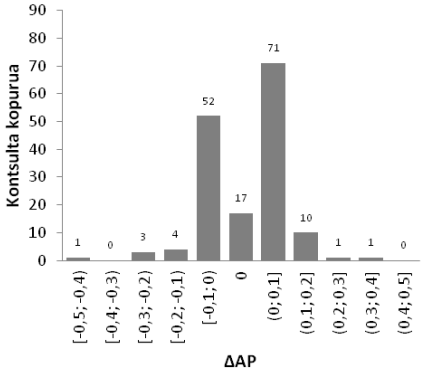
6.2 irudiak Robust datu-multzoko kontsulten konparaketetatik lortutakoak erakusten ditu. 6.2a, 6.2b eta 6.2c irudietan QL ereduarekiko konparaketak ageri dira. Emaitza orokorrekin bat datoz irudi hauek. Lehen ikusi dugu QLa baino hobeak direla beste hiru ereduak, eta lehen hiru irudi horietan QLa baino hobeto erantzundako kontsulten kopurua handiagoa dela



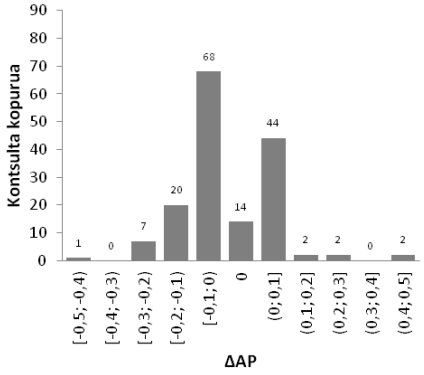
(a) PRFa QLarekiko



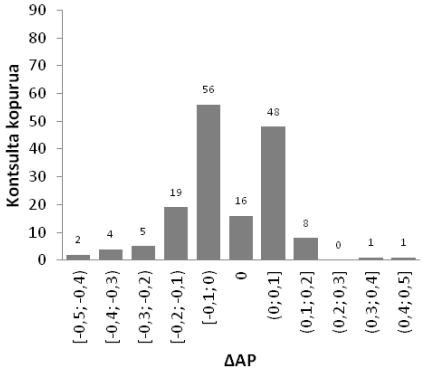
(b) RDEa QLarekiko



(c) RQEa QLarekiko



(d) RDEa PRFarekiko



(e) RQEa PRFarekiko

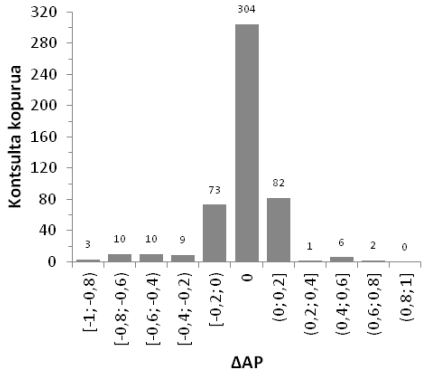
6.2 irudia – Robust datu-multzoko kontsultak hobekuntza-diferentzien arabera multzokatuak. (a), (b) eta (c) QL ereduarekiko hobekuntzak; (d) eta (e) PRF ereduarekiko hobekuntzak.

ikusten da (tarte positiboetako kontsulta kopurua handiagoa da negatiboetakoa baino). 6.2d eta 6.2e irudietan, berriz, hedapen-ereduak PRFarekiko konparatzen dira. Emaitzen taulan ikusi dugu (6.3 taula) PRFa guk proposatutako RDE eta RQE hedapen-eredua baino hobeto dabilela datu-multzo honetan. Hala ere, azken bi irudi hauetan ikus dezakegu hedapen-eredu hauek kontsulta batzuetan hobeto dabiltzala PRFa baino, kontsulta bat edo besteren batean 0,5erainoko AP diferentzia lortuz.

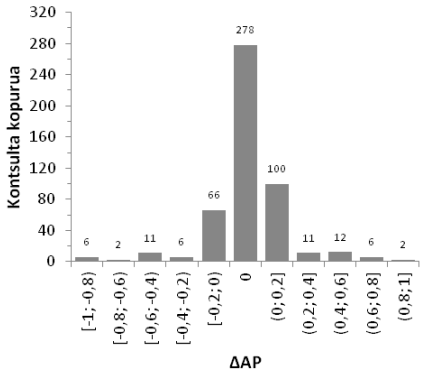
6.3 irudia ResPubliQA datu-multzoari dagokio. Irudietan ikusten den moduan kontsulten erdiak baino gehiagotan bi sistemen eraginkortasuna berdina izan da egindako konparaketa guztietan (gogoratu 500 kontsulta daudela datu-multzo honetan). 6.3a, 6.3b eta 6.3c irudietan QL ereduarekiko konparaketak ikus ditzakegu. Robustean ez bezala, QL eta PRFa konparatzen diren irudian ikusten da PRFarekin hobeto erantzundako kontsulten kopurua zerbait txikiagoa dela okerrago erantzundakoena baino. Berriz ere bat dator lehen ikusitako emaitza orokorrekin, ikusi baitugu PRFa ez dela eraginkorra datu-multzo honetarako. Aldiz, RDE eta RQE hedapen-ereduak QLa baino hobeto dabiltzala ikus daiteke berriz ere, hobeto erantzundako galderen kopurua handiagoa baita okerrago erantzundakoena baino. 6.3d eta 6.3e irudietan bi hedapen-ereduak PRFarekiko konparatzen dira, eta hauetan aldeak bai direla nabarmenak. Batez ere, RDE ereduak kontsulta gehiagotan hobeto egiten duela nabaria da.

Yahoo datu-multzoko konparaketak 6.4 irudian jarri ditugu. Irudi hauetan ere ikusten da datu-multzo honetako kontsulta askotan ez dagoela alderik edozein bi sistemen artean (0 diferentziako multzoa handia da beti). 6.4a, 6.4b eta 6.4c irudietan QL ereduarekiko konparaketak ikus ditzakegu. 6.4a irudian ikusten da PRFak ondo erantzuten dituen kontsulta kopurua handiagoa dela. Baina hobeto eta okerrago erantzundako kontsulta kopuruen arteko aldeak askoz handiagoak dira 6.4b (RDE) eta 6.4c (RQE) irudietan. Hori bat dator emaitzen taulan ikusitakoarekin, azken bi hauetan MRRaren hobekuntza handiagoa dela ikusi baitugu. 6.4d eta 6.4e irudietan bi hedapen-ereduak PRFarekiko konparatzen dira, eta hauetan ere RDE eta RQE ereduak hobeak direla argi ikusten da.

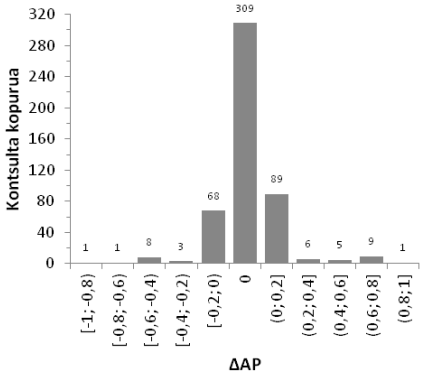
Bestetik, atal honetan proposatutako hedapen-ereduen portaera kontsulta zail eta errazei dagokienez ikusteko, beste konparazio hauek egin ditugu: ereduak binaka hartu eta trazatuak egin kontsulta bakoitzaren MAP (edo MRR) neurriaren arabera (6.5 irudia). Esaterako, 6.5a irudian Robust datu-multzoaren gainean PRF (ardatz bertikala) eta QL (ardatz horizontala) ereduen gauzatzea trazatu dugu. Dagokion joera-marrak esaten digu PRFak



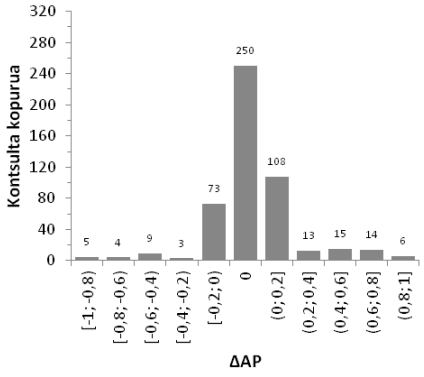
(a) PRFa QLarekiko



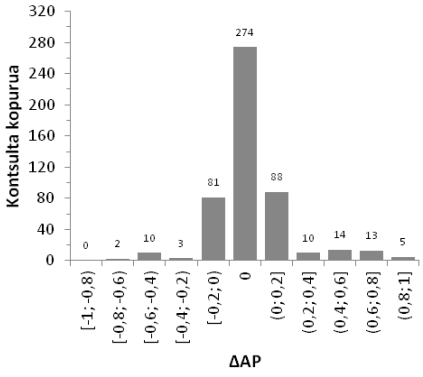
(b) RDEa QLarekiko



(c) RQEa QLarekiko

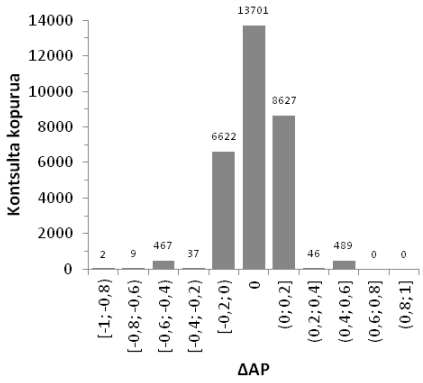


(d) RDEa PRFarekiko

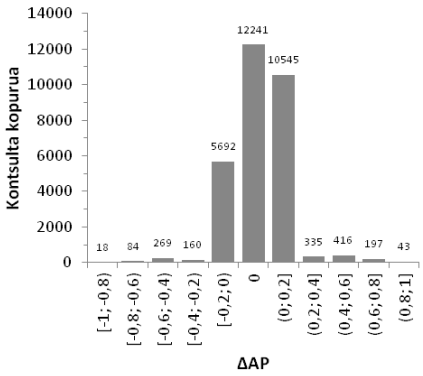


(e) RQEa PRFarekiko

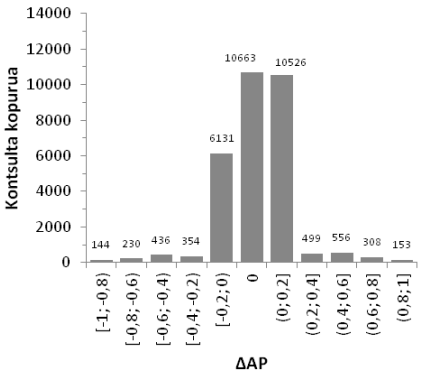
6.3 irudia – ResPubliQA datu-multzoko kontsultak hobekuntza-diferentzien arabera multzokatuak. (a), (b) eta (c) QL ereduarekiko hobekuntzak; (d) eta (e) PRF ereduarekiko hobekuntzak.



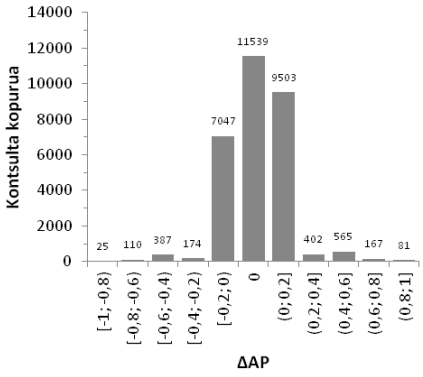
(a) PRFa QLarekiko



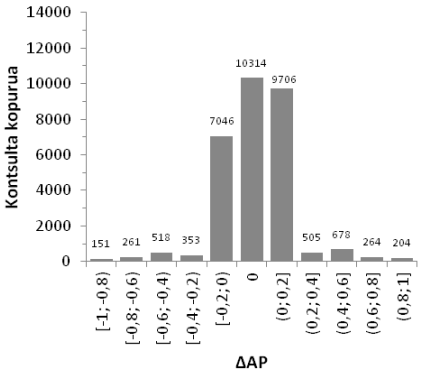
(b) RDEa QLarekiko



(c) RQEa QLarekiko



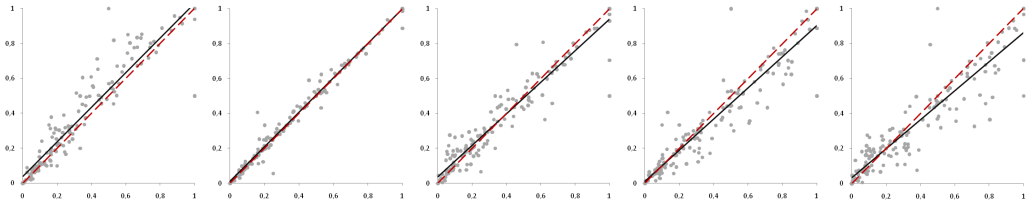
(d) RDEa PRFarekiko



(e) RQEa PRFarekiko

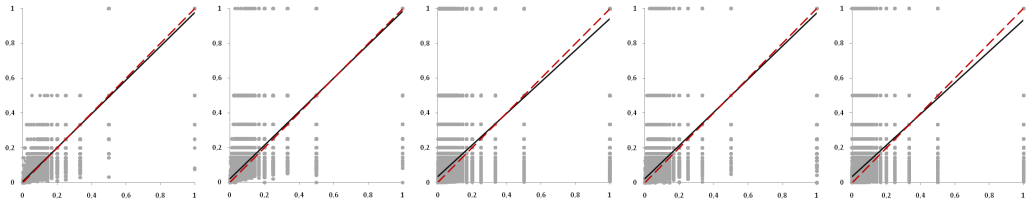
6.4 irudia – Yahoo datu-multzoko kontsultak hobekuntza-diferentzien arabera multzokatuak. (a), (b) eta (c) QL ereduarekiko hobekuntzak; (d) eta (e) PRF ereduarekiko hobekuntzak.

Robust



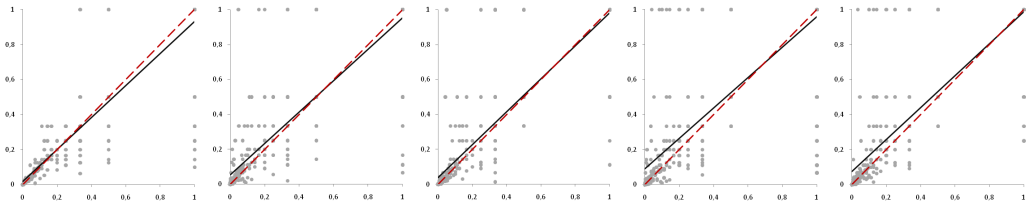
(a) QL-PRF **(b)** QL-RDE **(c)** QL-RQE **(d)** PRF-RDE **(e)** PRF-RQE

Yahoo!



(f) QL-PRF **(g)** QL-RDE **(h)** QL-RQE **(i)** PRF-RDE **(j)** PRF-RQE

ResPubliQA



(k) QL-PRF **(l)** QL-RDE **(m)** QL-RQE **(n)** PRF-RDE **(o)** PRF-RQE

6.5 irudia – Kontsulta guztien MAP edo MRR emaitzak. Lehen hiru zutabeetan PRF, RDE eta RQE hedapen-ereduak QL oinarri-lerroarekin alderatuz (QL X ardatzean). Laugarren eta bosgarren zutabeetan RDE eta RQE hedapen-ereduak PRF ereduarekin alderatuz (PRF X ardatzean). Dagozkien joera-marra linealak ere jarri ditugu (marra jarraitua).

QL hobetzen duela QLaren emaitza kontsulta jakin baterako edozein izanda ere. Beste bi bildumetan, aldiz, badirudi PRFa okerrago dabilela kontsulta errezetan, hots, QL oinarri-lerroak MAP altua lortzen duen kontsultetan (ikus 6.5f eta 6.5k irudiak).

RDE ereduaren kasuan (6.5b, 6.5g eta 6.5l irudiak), ikus dezakegu, kontsulta errazenetan QLaren portaera baino okerragoa bada ere, kontsulta zailenekin hobeto dabilela. RQEak ere antzeko portaera du (6.5c, 6.5h eta 6.5m irudiak).

PRFaren hobekuntza horiek RDE eta RQE ereduen hobekuntzen osagarriak direla esan daiteke. Hirugarren eta laugarren zutabeetan PRF vs RDE (6.5d, 6.5i eta 6.5n irudiak) eta PRF vs RQE (6.5e, 6.5j eta 6.5o irudiak) trazatuak, hurrenez hurren, jarri ditugu. Irudi horietan ageri diren joeramarrek adierazten dutenez, MAP altua duten kontsultentzat onuragarria da PRFa. Aldiz, kontsulta zailentzat, alegia, MAP baxua dutenentzat, RDE eta RQE ereduak dira onuragarrienak.

Kontsulten banakako azterketa horretatik adibide batzuk atera ditugu. Esaterako, sarrerako kapituluko 1.2 irudian ikusi ditugun adibideak zailak direla esan dezakegu, eredu guztien (QL, PRF, RDE eta RQE) MRR emaitza 0 izan baita bi kontsulta horientzat. Beste adibide batekin jarraituz, kapitulu honetan ahaidetasunaren bidez kontsulta baten hedapena nola egiten den erakusteko erabili dugun ResPubliQAko adibide horren kasuan, QL, PRF eta RDE ereduek 0,3333ko MRR emaitza lortu dute. Kontsultaren hedapena eginez (RQE eredua erabiliz), aldiz, MRR=1 lortu dugu, alegia, dokumentu adierazgarria lehenengo postuan itzuli dugu. 6.6 irudian jarri dugu, berriz ere, kontsulta hori bera (6.6a), horren hedapenetik lortutako hitz batzuk (6.6b) eta kontsulta horrentzako dokumentu adierazgarria (6.6c). Hedapena eginez dokumentu adierazgarri hori topatzeko arrazoietako bat hedapenean *vehicle* eta *distance* hitzak lortu ditugula izan daiteke, hitz horiek dokumentuan ere agertzen baitira. Gainera, hedapenean *miles per hour* terminoaren sinonimo moduan *mph* ere lortzen dugu, eta dokumentuan ere sinonimo biak daude.

6.5.3 Parametroen eraginaren analisia, ereduak konparatuz

6.1 taulan erakutsi ditugu metodo eta datu-multzo bakoitzerako parametrorik hoberenak. Errealitatean, datu-multzo berri batekin lanean hasi behar

What is the lowest speed in miles per hour which can be shown on a speedometer?

(a) Ingelesezko 91. galdera.

speedometer speed_indicator miles_per_hour **mph** motor_vehicle automotive_vehicle low read register show record **vehicle** speed velocity display show **distance** length

(b) Kontsultaren hedapenetik lortutako hitz batzuk.

where a **vehicle** is intended for sale in a Member State where imperial **distances** are used, the speedometer must also be graduated in **mph** (miles per hour), with subdivisions of 1, 2, 5 or 10 mph. Marked numerical speed value intervals must not exceed 20 mph and must begin at either 10 mph or 20 mph;

(c) Dokumentu adierazgarria (jrc32000L0007-en/92).

6.6 irudia – ResPubliQA datu-multzoko galdera bat, kontsulta-hedapenaren bidez zuzen erantzun ahal izan duguna.

dugunean, askotan ez da egoten datu-multzo horren gainean parametroen doitzea egiteko entrenamendurako datu-multzorik. Horrelakoetan, beste ingurune batzuetarako hoberenak diren parametroen balioak erabiltzea da aukera bakarra. μ leuntze-parametroak lotura zuzena du dokumentuaren luzerarekin, eta, hortaz, beste esperimentu batzuetan oinarrituta nahiko erraz finkatu daiteke honen balioa.

Honela bada, beste bildumetan optimizatutako parametroekin ere egin ditugu esperimentuak (μ ez dugu aldatu, datu-multzo horretako baliorik optimoena emango diogu). 6.4 taulan esperimentu hauen eta datu-multzoan bertan optimizatutako parametroekin egindako esperimentuen arteko emaitzen aldeak jarri ditugu. Azken lerroko batez besteko balioak erakusten digu RDE eredua dela optimizaziorekiko sentikortasun gutxiena duena. RDE ereduan emaitzak hobetzen dira, RQE ereduan zerbait okerragoak dira emaitzak, eta galera handienak PRF ereduan ditugu.

Ondoren, eredu bakoitzean parametro bakoitzak duen eragina aztertuko dugu. Horretarako, aztertzen ari garen parametro horren balio desberdinetarako sistemak lortzen dituen emaitzak irudikatuko ditugu, beti ere beste parametroentzat beraien balio optimoenak finkatuz. Azterketa hauek parametroen balio horiek optimizatzeko erabilitako entrenamenduko datu-multzoaren gainean egin dira. Hortaz, kontuan izan, orain ikusiko ditugun irudietako emaitza horiek ez direla test ataleko emaitzak.

	PRF			RDE			RQE		
datu-multzoa	Rob	Yah	Res	Rob	Yah	Res	Rob	Yah	Res
Robust	—	-9,7	-18,8	—	0,0	1,0	—	-4,3	-7,6
Yahoo!	-6,3	—	1,7	0,0	—	1,0	0,5	—	-0,4
ResPubliQA	-7,3	-0,9	—	-0,7	-0,7	—	0,9	1,3	—
batez beste	% -6,9			% 0,11			% -1,60		

6.4 taula – Esperimentu optimo eta beste bildumetan optimizatutako parametroak erabiliz egindako esperimentuen emaitzen arteko alde erlatiboak (MAP edo MRR). Lehenengo zutabean parametroak zein bildumetan optimizatu diren zehazten da. Azken lerroko balioak eredu horretako diferentzien batez bestekoak dira.

Termino kopurua

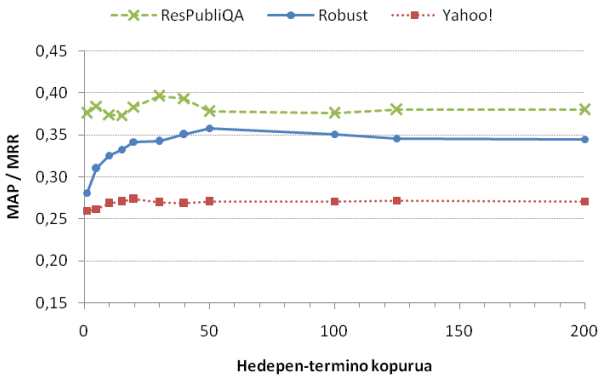
Kontsulta-hedapena egiterakoan optimizatu beharreko parametroetako bat hedapenean erabili beharreko termino kopurua da. 6.7 irudian ikus daiteke PRF eta RQE ereduen portaera termino kopuruarekiko. 6.7a irudian ikusten denez, PRF ereduak termino kopurua aldatzean gorabeherak ditu. Aldiz, 6.7b irudiak erakusten du RQE ereduan termino kopurua handitzen den hein berean, emaitzak gorantz egiten duela, harik eta goi-muga batera iristen den arte. Portaera hau datu-multzo guztietan antzekoa da. PRFaren kasuan 20-50 tartean dago termino kopuru optimoena. RQE ereduak termino gehiago behar ditu hobeto ibiltzeko, 50-125 tartean baitago kontzeptu kopururik egokiena datu-multzoaren arabera. Kontuan izan, RQEaren kasuan, kontzeptu kopurua dela parametroan finkatzen dena, eta ez termino kopurua.

Dokumentu kopurua

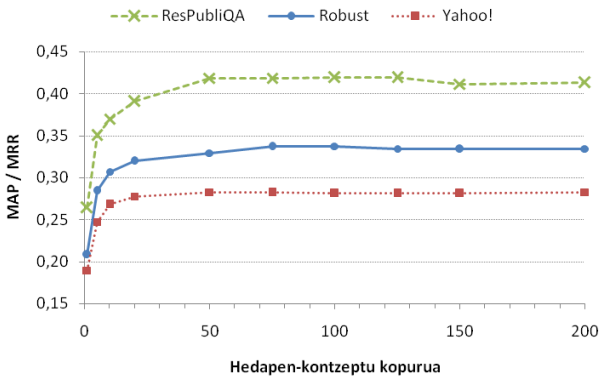
PRF ereduak bakarrik du dokumentu kopurua zehazteko parametroa, eta 6.8 irudian ikus daiteke eredu horretan datu-multzo desberdinen portaera. Eredu bakarra izanda, ezin da parametro honekiko eredu desberdinek duten portaeraren konparaketarik egin.

Jatorrizko kontsultaren pisua

6.9 irudian hemen aurkeztutako eredu bakoitzean jatorrizko kontsulten pisuaren araberako emaitzak (MAP edo MRR) ageri dira. PRFari dagokio-



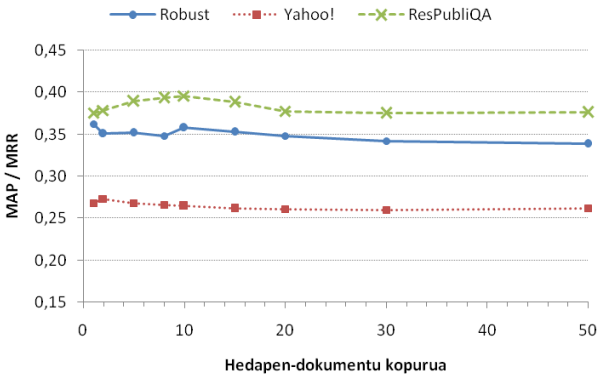
(a) PRF



(b) RQE

6.7 irudia – Hedapen-termino kopuruaren araberako emaitzak PRF eta RQE ereduetarako.

nez (6.9a), datu-multzo bakoitzaren portaera nahiko desberdina da. Robust da jatorrizko kontsultari pisu gutxien ematen diona (0,3an dago maximoa). Beste bi bildumek 0,8an lortzen dute emaitzarik hoberena. Eta Yahoo! da pisu honek gutxien eragiten dion datu-multzoa. RDE ereduan hiru bildumek portaera nahiko antzekoa dute, emaitzarik altuena 0,7-0,8 tartean dutelarik. Aipatzekoa da, eredu honetan hedapeneko terminoak bakarrik erabiliz entrenamenduko datu-multzoan lortutako emaitzak (pisua 0 denean) oso kaxkarrak badira ere beste ereduekin alderatuz, balio optimoenak erabiliz, *test* datu-multzoan beste ereduekin lortutako emaitzetara inguratzen dela, eta, esaterako, Yahoo! datu-multzoan RDE eredu honekin lortzen dela emaitza-



6.8 irudia – Hedapenean erabilitako dokumentu kopuruaren araberako emaitzak PRF eredurako.

rik hoberena. RQE ereduan ere, Yahoo! eta ResPubliQA bildumen portaera oso antzekoa da, 0,7an lortuz beraien emaitzarik hoberena. Robust datu-multzoan jatorrizko eta hedapeneko kontsultei pisu berbera emanez lortzen da emaitzarik hoberena.

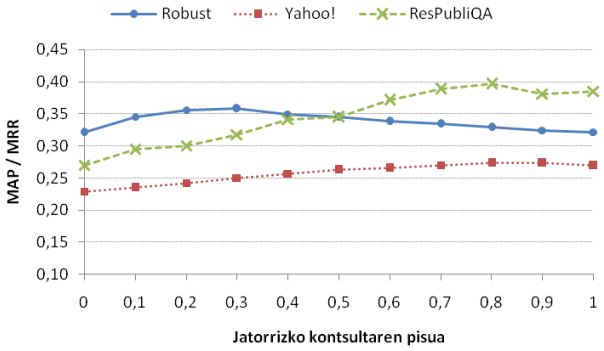
6.5.4 Parametroen eraginaren analisia, datu-multzoak konparatuz

Atal honen hasieran sistemen emaitza orokorrak ikusi ditugunean esan dugu datu-multzo batetik bestera eredu bakoitzak portaera desberdina duela. Alegia, ez dago eredu bat datu-multzo guztietarako onena denik. Honen atzean hainbat arrazoi egon daitezke. Esaterako, Xu eta Croft-ek (2000) diotenez, PRFa hobeto dabil kontsulta motzak dituzten bildumetan. Hain zuzen ere, gure kasuan horixe gertatzen da. Izan ere, Robust datu-multzoko kontsultak motzagoak dira Yahoo! eta ResPubliQA bildumetakoak baino, eta PRF ereduarekin emaitzarik hoberenak datu-multzo horretan bertan lortzen ditugu.

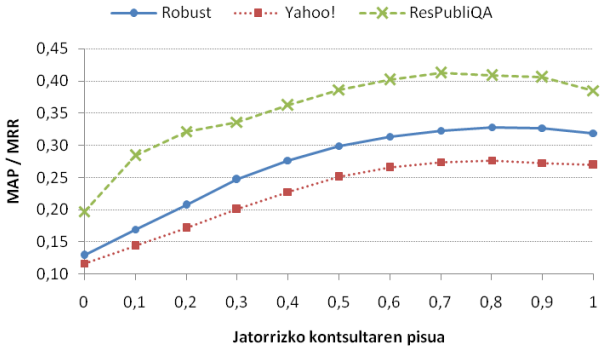
Baina beste hainbat faktore ere egon daitezke portaera hori justifikatu dezaketenak. Hori dela eta, datu-multzo bakoitzean eredu desberdinetan parametroek duten eragina aztertuko dugu jarraian.

Termino kopurua

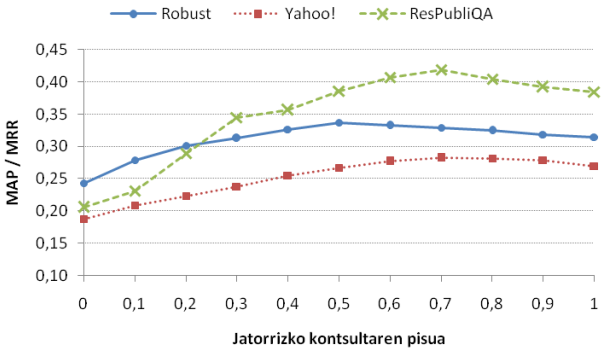
PRF ereduko parametroarekin termino kopurua zehazten da, eta RQE eredukoarekin kontzeptu kopurua. Bi kopuruek gauza desberdinak neurtzen di-



(a) PRF



(b) RDE



(c) RQE

6.9 irudia – Jatorrizko kontsultaren pisuaren araberako emaitzak eredu bakoitzerako

tuztenez, ez dira konparagarriak eta ezin ditugu irudi edo grafiko berdinean konparatu eredu desberdinak.

Dokumentu kopurua

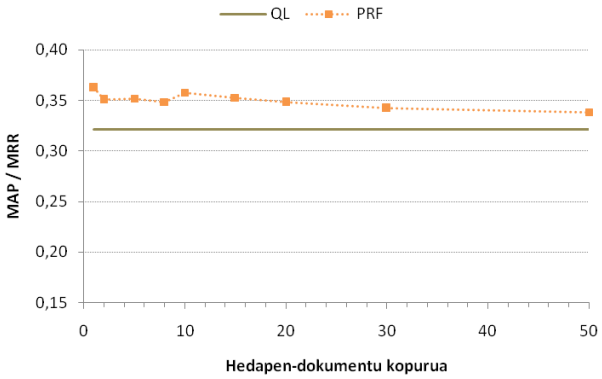
PRFan galerak gertatzearen arrazoietako bat lehenengo itzuliko dokumentu zerrendaren buruan dauden dokumentuak adierazgarriak ez izatea izan daiteke. Fenomeno honi ingelesez *topic drift* deitzen zaio (Mitra *et al.*, 1998). Gure PRF esperimentuetan gertatutakoa *topic drift* delakoaren ondorioz izan daiteke neurri batean. Izan ere, Robust datu-multzoan gai bakoitzarentzat dokumentu egoki dezente dauden bitartean (dokumentu guztiak albistetakoak dira, eta, beraz, PRFrako egokia da datu-multzo hau), Yahoo! datu-multzokoak elkarren artean zerikusirik ez duten kontsultei erantzuten dieten dokumentuak dira. Horregatik, *topic drift* gertatzeko arrisku gehiago dago aipatutako azken datu-multzo honetan. Gauzak horrela, Yahoo! datu-multzoaren kasuan hedapenean erabiltzeko dokumentu kopururik hoberena 2 da, eta Robust datu-multzoan, berriz, 10. Horixe ikusten da 6.10 irudietan. Yahoo! datu-multzoaren kasuan, egindako probetan 2 dokumenturekin bakarrik lortzen du QL hobetzea. Robust datu-multzoaren kasuan, aldiz, edozein dokumentu kopuru erabilita ere, beti QLa baino hobeto dabil PRF eredua.

Jatorrizko kontsultaren pisua

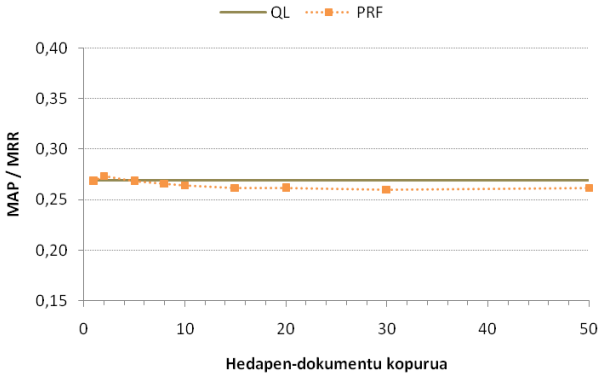
6.11 irudian datu-multzo bakoitzean hedapeneko kontsultari emandako pisuak duen eragina aztertu dezakegu. Robust datu-multzoan (6.11a irudia) jatorrizko kontsultaren pisu optimo txikiena duena PRFa da (0,3 balioarekin lortzen du emaitzarik onena). Hala ere, PRF ereduak jatorrizko kontsultari edozein pisu emanda ere, ia beti hobetzen du QL oinarri-lerroa. Ez da horrelakorik gertatzen beste bi bildumetan (ikus 6.11b eta 6.11c irudiak). Hauetan hiru ereduen portaera nahiko antzekoa da, jatorrizko kontsultaren pisua 0,7 edo 0,8 baino handiagoa denean bakarrik lortzen dute QL oinarri-lerroa hobetzea.

6.5.5 Ereduak konbinatzeko aurretiazko esperimentuak

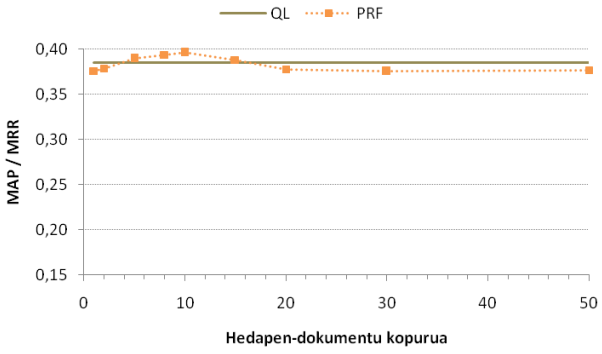
Egiten ari garen emaitzen analisi honetan ikusi da guk proposatutako metodoak PRF ereduarekiko osagarriak direla. Hori dela eta, metodo hauen



(a) Robust

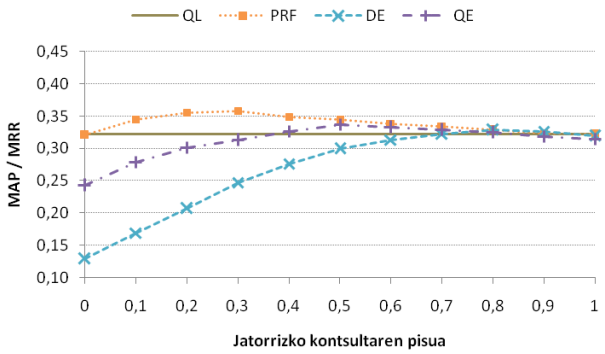


(b) Yahoo!

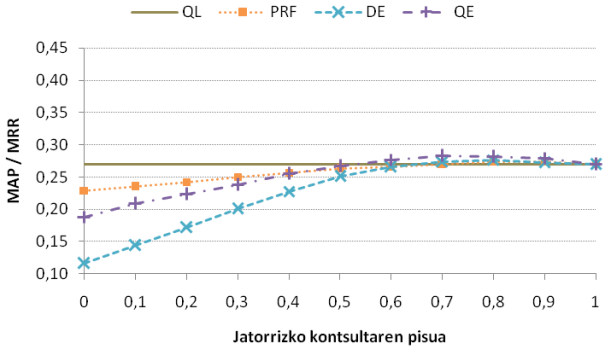


(c) ResPubliQA

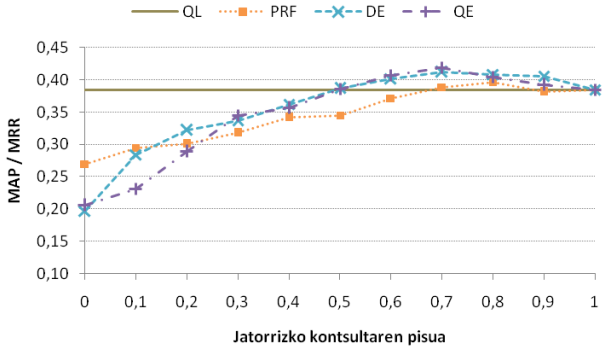
6.10 irudia – Hedapenean erabilitako dokumentu kopuruaren araberako emaitzak datu-multzo bakoitzerako.



(a) Robust



(b) Yahoo!



(c) ResPubliQA

6.11 irudia – Jatorrizko kontsultaren pisuaren araberako emaitzak datumultzo bakoitzerako.

konbinazioa hurbilpen egoki bat izan zitekeela pentsatuz, PRF eta RQE konbinatuz aurretiazko esperimentu bat egin dugu Robust datu-multzoaren gainean. Esperimentu honetan, kontsultaren hedapenerako PRF ereduan lortutako hedapen-terminoak eta RQE eredutik lortutako hedapen-terminoak erabili ditugu, guztiak batera. Hedapena modu honetan eginda MAP neurrian 0,3767ko emaitza lortu dugu, eta 0,1543 GMAP neurrian. Hortaz, orain arteko emaitzarik onenak eredu konbinatu honekin lortu ditugu.

6.6 Ondorioak

Kapitulu honetan aurreko kapituluan aurkeztutako hedapen-teknika bera erabili dugu. Teknika WordNeten oinarritutako grafo-algoritmo baten bidez kontsulta edo dokumentuarekin erlazionatutako kontzeptuak lortzean datza. Behin kontzeptuak lortuta, hauei dagozkien terminoekin kontsultak eta dokumentuak hedatu ditugu. Kontsulten hedapenean kontsulta egituratuak erabili ditugu; dokumentuen hedapenean, berriz, bi indize erabili ditugu: bat jatorrizko hitzentzat eta bestea hedapenetik eratorritako hitzentzat.

Aurreko kapituluko esperimentuetatik hauetara dagoen aldaketarik handiena erabilitako IB sisteman dago, oraingo honetan hizkuntza-ereduetan oinarritutako IB sistema batean txertatu baitugu hedapen-teknika. Gainera, guk proposatutako kontsulta eta dokumentuen hedapen-ereduak (RQE eta RDE, hurrenez hurren) oinarri-lerro arrunt batekin konparatzeaz gain, oso erabilia eta emaitza onak lortzen dituen *pseudo-relevance feedback* (PRF) ereduarekin ere konparatu ditugu.

Aurreko ataletan ikusitakoaren arabera, ondorengo ikerketa-galdera hauek erantzun ahal izan ditugu:

– **IG 2** – ***Ezagutza-base lexikal bidezko ahaidetasun semantikoan oinarritutako hedapena eginez hobetzen al da IB sistemaren eraginkortasuna?***

WordNeten oinarritutako grafo-algoritmo baten bidez kontsulta eta dokumentuak semantikoki erlazionatutako hitzekin hedatu ditugu, kontsulten hedapenerako kontsulta egituratuak eta dokumentuen hedapenerako bi indize erabiliz. Hedapen-eredu hauek, oraingo honetan, hizkuntza-ereduetan oinarritutako IB sistema batean txertatu ditugu. Izaera desberdineko hiru datu-multzotan esperimentuak eginez, hemen proposatutako bi hedapen-ereduekin

hobekuntzak lortu ditugu *query likelihood* oinarri-lerroko sistemarekiko datu-multzo guztietan.

- 2.2 - *Hedapen-teknika hau egokia al da kontsulta- zein dokumentu-hedapenerako? Bi hauetakoren bat ba al da bestea baino eraginkorragoa?*

  Kapitulu honetako esperimentuetan hedapen-teknika hau kontsultak eta dokumentuak hedatzeko erabili dugu. Hedapena egiteko modua berdina da kontsultentzako eta dokumentuentzako. Desberdintasun bakarra hedapenetik eratorritako hitz berri horiek IB sisteman txertatzeko moduan dago. Horretarako, kontsulten kasuan 6.3.2 atalean aurkeztutako RQE eredua erabili dugu. Eredu horretan jatorrizko kontsultako hitzekin eta hedapenetik lortutako hitzekin kontsulta egituratu bat sortzen dugu. Dokumentuen kasuan, berriz, 6.3.1 atalean aurkeztutako RDE ereduan bi indize sortu (bat jatorrizko hitzekin eta bestea hedapeneko hitzekin) eta berauek konbinatzea proposatzen dugu. Hizkuntza-ereduetan oinarritutako bi hedapen-eredu hauetatik eraginkorrena zein den ezin dezakegu esan, datu-multzoen artean emaitzak kontrajarriak baitira: Yahoo! datu-multzorako bi ereduekin hobekuntza nabarmenak lortzen dira, baina, bietatik bat aukeratzekotan, RDE eraginkorragoa da; Robust datu-multzorako ere, oro har, RDE ereduarekin emaitza hobeak lortzen dira; aitzitik, ResPubliQA datu-multzorako egokiena RQE eredua dela esan daiteke.

- 2.3 - *Hedapen-teknika hau pseudo-relevance feedback metodoarekin alderatzean, zer ikusten dugu?*

  RQE eta RDE ereduekin lortutako emaitzak PRF (*pseudo-relevance feedback*) ereduarekin lortutako emaitzekin konparatuz ateratzen ditugun ondorioak desberdinak dira datu-multzoaren arabera: RDE eta RQE dira eraginkorrenak Yahoo! eta ResPubliQA bildumetarako; Robust datu-multzorako, ordea, ez. Hala ere, kontsulta bakoitzaren emaitzak banan-banan aztertuz gero, horietako batzuetarako gure hedapen-ereduak PRF eredua baino eraginkorragoak direla ikusi dugu. Emaitzen analisiek hedapen-ereduak eta PRF eredua osagarriak direla erakutsi digute, PRF eredua galdera errazentzat hobea eta gure hedapen-ereduak galdera zailenetan eraginkorragoak baitira. Are gehiago, Robust datu-

multzoan RQE eredua PRF ereduaren pareko dagoela adierazten digute GMAP balioek.

- 2.4 - *Hedapen-teknika honen eraginkortasunean zein faktorek eduki dezakete eragina?*

Hedapen-teknika honen sendotasunean eta eraginkortasunean parametroek eduki dezaketen eragina aztertu nahi izan dugu. Hurrengo zerrenda honetan analisi horretatik ateratako ondorioak zein izan diren ikus ditzakegu:

- Parametroak orokorrean.
  Datu-multzo berri batekin lanean hasi behar dugunean, askotan ez da egoten datu-multzo horren gainean parametroen doitzea egiteko entrenamendurako datu-multzorik. Horrelakoetan, aukera bakarra, beste ingurune batzuetarako hoberenak diren parametroen balioak erabiltzea da. Horrelako kasu batean aurrean egongo bagina, guk proposatutako hedapen-ereduen eraginkortasuna nola aldatuko litzatekeen aztertu dugu eta RDE eredua optimizazioarekiko sentikortasun gutxiena duena dela ikusi dugu, eredu honen emaitzetan gorabehera gutxi baitaude parametroak aldatu arren.
- Hedatu beharreko termino edo kontzeptu kopurua.
  RQE ereduan kontsultaren hedapeneko termino kopurua handitzen den heinean emaitzek gorantz egiten dute, harik eta goi-muga batera iristen den arte. Datu-multzoaren arabera, 50-125 tartean dago kontzeptu kopururik egokiena. Esan behar da, PRF ereduak termino kopurua aldatzean gorabehera gehiago baditu ere, kopuru txikiagoekin ondo moldatzen dela (20-50 tartean termino kopurua), eta hori sistemaren errendimenduari begira, ezaugarri positibo bat da. RDE ereduan ez dugu probarik egin dokumentua hedatzerakoan erabilitako kontzeptu kopurua aldatzeak duen kostuagatik eta aurreko kapituluetan antzeko esperimentuetan parametro honen eragina hutsala zela ikusi genuelako; beti 100 kontzeptu hedatu ditugu.
- Jatorrizko kontsultaren pisua.
  RDE ereduan hiru datu-multzoek antzeko portaera dute parametro hau aldatzean, eta parametro honek 0,7-0,8 tarteko ba-

lioak hartzean lortzen dira emaitzarik onenak. RQE ereduan tartea 0,5-0,7an dago. Horrek esan nahi du, hedapen-eredu hauetan ia beti jatorrizko hitzei garrantzi handiagoa emanez lortzen direla emaitzarik onenak. PRF ereduan, aldiz, ez da beti hori gertatzen, Robust datu-multzorako pisurik onena 0,3 delako; alegia, hedapeneko hitzei askoz indar handiagoa esleitzen die. Hala ere, esan daiteke Robusten portaera PRF ereduaren kasuan salbuespena dela, beste datu-multzoek eredu guztietan, baita PRF ereduan ere, portaera antzekoa dute: pisua 0,7 edo 0,8 baino altuagoa denean bakarrik hobetzen dute QL oinarri-lerroko sistema.

Laburbilduz, RQE eta RDE ereduek galdera zailetan duten eraginkortasuna ikusirik, ahalmen hori gehiago lantzea nahiko genuke etorkizunean, beharbada PRFarekin konbinatuz. Bestetik, hedapenetik lortzen dugun informazio guztia beste nolabait ustiatzeko aukerak landu nahiko genituzke; adibidez, Mei *et al.*-en (2008) eta Huang *et al.*-en (2009) lanetan oinarrituz, hizkuntza-ereduen leuntzea egiteko erabil genezake hedapeneko informazio hori.

7

Ondorioak eta etorkizuneko lanak

Lan hau aurkezterakoan ikusi dugu hizkuntzaren fenomeno linguistiko batzuk IBrako arazo-iturri direla. Arazo horietako batzuei irtenbide nahiko egokia eman zaie hizkuntzaren prozesamenduko (HP) hainbat teknika erabiliz; esate baterako, bariazio morfosintaktikoek sorrarazten dituzten zailtasunak lematizatzailea erabiliz gaindi daitezke, neurri batean behintzat. Semantika tarteko duten arazoak, ordea, ez dira ebazteko hain samurrak. Arazoa teorikoki aztertuz gero, badirudi hizkuntzaren prozesamendurako hainbat teknikek ahalbidetu dezaketela arazo horiek konpontzea. Saiakera eta ikerketa-lan asko egin izan dira hizkuntzaren prozesamendurako teknikak erabiliz sinonimiaren eta polisemiaren eragin kaltegarriak ekidin nahian, horietako batzuetan emaitza itxaropentsuak lortu izan direlarik. Hori ikusirik, guk ere bide horretatik jarraitu eta tesi-lan hau aurrera eraman dugu, sarrerako atalean mahaigaineratu dugun galdera nagusi honi erantzuna bilatu nahian:

> **Kontsulten eta dokumentuen hedapenerako semantika lexikala erabiliz hobetzen al da IB sistemen eraginkortasuna ad hoc atazetan?**

Amaierako kapitulu honetan, lehenik, galdera horri eta beste ikerketa-galderei aurreko kapituluetan emandako erantzunak bilduko ditugu. Jarraian, tesi-lan honen bidez egindako ekarpen nagusiak zeintzuk diren laburbilduko dugu. Amaitzeko, etorkizuneko lanen berri emango dugu.

7.1 Ikerketa-galderen erantzunak

Ikusi berri dugun galdera nagusi horretatik abiatuz, beste hainbat ikerketa-galdera zehatzago proposatu genituen (ikus 1.5 atala). Esperimentuak azaldu ditugun aurreko kapituluetan galdera horiei erantzun diegu, eta jarraian, galdera horiek guztiak beren erantzunekin hona ekarriko ditugu.

– **IG 1** – ***Hitzen adiera-desanbiguazioan eta ezagutza-base lexikal bateko sinonimoetan oinarritutako hedapena eginez hobetzen al da IB sistemaren eraginkortasuna?***

4. kapituluko esperimentuetan HAD informazioarekin etiketatutako gai- eta dokumentu-bildumez baliatuz, kontsulta- eta dokumentu-hedapena egin ditugu WordNeteko sinonimoak erabiliz. Baliabide horiekin eta parametroak doitu gabe ataza elebakarrean (ingelesa) berreskurapeneko emaitzak hobetzea lortu dugu oinarrilerroko sistemarekiko, nahiz eta lortutako hobekuntza ez izan estatistikoki esanguratsua.

- 1.1 - *Hedapen-teknika hau egokia al da kontsulta- zein dokumentu-hedapenerako? Bi hauetakoren bat ba al da bestea baino eraginkorragoa?*

Gure esperimentuetan bai kontsultak bai dokumentuak hedatu ditugu. Erabiltzaileak egin ohi dituen kontsultak nahiko motzak izan ohi dira, eta, hortaz, gerta daiteke ez edukitzea nahiko testuinguru adiera-desanbiguazioa zuzen burutzeko. Baina, erabili dugun datu-multzoko gaien *title* eta *description* eremuak erabili ditugunez, esperimentu hauetako kontsultak nahiko luzeak dira. Gainerakoan, behin HAD informazioa izanda, hedapen-prozesua berdin-berdina da kontsultentzako ala dokumentuentzako. Desberdintasuna hedapenetik lortutako hitz berri horiek IB sisteman txertatzeko moduan dago. Hainbat kontsulta egituratu konplexurekin esperimentuak egin ditugu, jatorrizko hitzak eta hedapenetik lortutakoak konbinatzeko kontsultarik eraginkorrena zein den jakiteko. Baina, emaitzarik hoberenak dokumentuen hedapena bakarrik eginez lortu ditugu.

- 1.2 - *Hedapen-teknika honen eraginkortasunean zein faktorek eduki dezakete eragina?*

Hedapen-teknika honen hainbat aldaerarekin egin ditugu probak, eta hauek dira esperimentu horietatik ateratako ondorioetako batzuk:

- Hedapen-mota: osoa (hitz bakoitzaren adiera guztien sinonimo guztietara hedatu) vs onena (hitz bakoitzaren pisurik handieneko adieraren sinonimo guztietara hedatu).
  Kontsulten hedapenean, orokorrean esanda, erabateko hedapenarekin emaitza hobeak lortzen dira; dokumentuen hedapenean, aldiz, hedapen onena da eraginkorrena.
- Kontsultaren luzera, kontsulta osatzeko erabiliko diren gaiaren eremuak: *title* vs *title+description*.
  Esperimentu nagusienetan *title+description* erabili dugu, eta horrela hedapena eginez estatistikoki esanguratsuak ez diren hobekuntza txikiak lortu ditugu. *title* bakarrik erabiliz ea zer gertatzen zen ikusi nahi izan dugu. Proba horietatik ondorio garbirik ezin izan dugu atera: entrenamendurako bildumarekin ez dugu hobekuntzarik lortu, baina, testeko fasean hobekuntza handia, esanguratsua dena, lortu dugu. Hortaz, ezin esan ziurtasun osoz kontsulta motzetan (2-3 hitz) hedapenak eragin handiagoa duenik.
- Kontsulta eta indizeetan erabilitako unitatea: lema vs *synset*a.
  Hedapena egitean hitz berriak sartzen dira IB prozesuan, eta, hortaz, zarata sartzeko arriskua dago, hedapen desegokia egiteagatik edo gehitutako hitz berriak polisemikoak direlako. Hori ekiditeko, hedapenean lemak erabili beharrean, *synset*ak erabiliz proba batzuk egin ditugu. *Synset*ak ez direla erabilgarriak ikusi dugu esperimentu hauetan.
- HAD sistema desberdinak: UBC vs NUS.
  Bi HAD sistema desberdinen irteerekin probak egin ditugu, baina ezin izan dugu ondorio garbirik atera, esperimentu batzuetan sistema batekin lortu baititugu emaitzarik hoberenak, eta alderantziz. Hala ere, kontaketak eginez gero, NUS sistema gailentzen dela esan daiteke. Aipatu behar da, SemEval-2007ko *all-words WSD* atazan NUS sistemak UBC sistemak baino emaitza zertxobait hobeak lortu zituela (Pradhan *et al.*, 2007).

Esperimentu hauetan erabilitako IB sistemak hainbat parametro ditu, hala nola, leuntze-teknikarena, PRF teknikarenak eta hedapeneko kontsultaren pisua. Parametro hauen balio desberdinekin hainbat konbinaketa probatu baditugu ere, ez dugu balio optimoaren zantzu garbirik atera.

- 1.3 - *Hedapen-teknika hau egokia al da kontsulten eta dokumentuen itzulpena egiteko hizkuntza arteko berreskurapenean?*

WordNet hainbat hizkuntzatarako dagoenez, kontzeptu baten *synset* zenbakia zein den jakinda, oso erraz lortu daitezke kontzeptu horri dagozkion hitzak hainbat hizkuntzatan. Hitz horiek jatorrizko hizkuntzakoak izan beharrean beste hizkuntza bateko WordNetetik hartzen baditugu, itzulpena egiten ariko gara, hedapenaz gain. Hortaz, hedapen-teknika honek kontsulta eta dokumentuak itzultzeko balio du. Itzulpen-metodo hau hizkuntza arteko ataza batean (gaztelania-ingelesa) erabili eta berreskurapeneko emaitzak hobetzea lortu dugu, estatistikoki esanguratsuak diren hobekuntzak, gainera.

– **IG 2** – ***Ezagutza-base lexikal bidezko ahaidetasun semantikoan oinarritutako hedapena eginez hobetzen al da IB sistemaren eraginkortasuna?***

5. eta 6. kapituluetan WordNeten oinarritutako grafo-algoritmo baten bidez kontsultarekin (edo dokumentu bakoitzarekin) semantikoki erlazionatutako kontzeptuak lortu eta ahaidetasun handieneko kontzeptuak lexikalizatzen dituzten hitzekin kontsulta (edo dokumentua) hedatzen dugu. Kontsulta eta dokumentu hedatuak erabiliz hainbat berreskurapen-atazen eraginkortasuna hobetzen dela ikusi dugu. Esperimentuak hainbat dokumentu-bilduma, parametro-ezarpen desberdinekin eta beste hainbat aldaerekin egin ditugu, orokorrean emaitza positiboak lortuz.

- 2.1 - *Hedapen-teknika hau eraginkorra al da izaera desberdineko berreskurapen-ereduetarako?*

Ahaidetasunean oinarritutako hedapen-teknika hau bi motatako IB sistematan txertatu dugu: berreskurapen-eredu probabilistiko klasikoan (5. kapitulua) eta hizkuntza-ereduetan oinarritutako

berreskurapen-ereduan (6. kapitulua). Oro har, biekin emaitza positiboak lortu ditugu.

- 2.2 - *Hedapen-teknika hau egokia al da kontsulta- zein dokumentu-hedapenerako? Bi hauetakoren bat ba al da bestea baino eraginkorragoa?*

  Ahaidetasunean oinarritutako hedapen-teknika honek edonolako testuak hedatzeko balio du, hedapen-prozesua beti berdina delarik. Guk bai kontsultak eta bai dokumentuak hedatzeko erabili dugu. Kontsulta- edo dokumentu-hedapena, bi hauetan bat bestea baino eraginkorragoa izatea dokumentu-bildumaren menpe dagoela ikusi dugu, hizkuntza-ereduetan oinarritutako berreskurapen-eredua darabilgunean behintzat (6. kapitulua).

- 2.3 - *Hedapen-teknika hau pseudo-relevance feedback metodoarekin alderatzean, zer ikusten dugu?*

  Hizkuntza-ereduetan oinarritutako berreskurapen-eredua darabilgunean RQE (kontsulta-hedapena) eta RDE (dokumentu-hedapena) ereduekin lortutako emaitzak PRF (*pseudo-relevance feedback*) ereduarekin lortutako emaitzekin konparatuz ateratzen ditugun ondorioak desberdinak dira datu-multzoaren arabera: RDE eta RQE dira eraginkorrenak Yahoo! eta ResPubliQA bildumetarako; Robust datu-multzorako, ordea, ez. Hala ere, kontsulta bakoitzaren emaitzak banan-banan aztertuz gero, horietako batzuetarako gure hedapen-ereduak PRF eredua baino eraginkorragoak direla ikusi dugu. Emaitzen analisiek hedapen-ereduak eta PRF eredua osagarriak direla erakutsi digute, PRF eredua galdera errazentzat hobea eta gure hedapen-ereduak galdera zailenetan eraginkorragoak baitira. Are gehiago, Robust datu-multzoan RQE eredua PRF ereduaren pareko dagoela adierazten digute GMAP balioek.

- 2.4 - *Hedapen-teknika honen eraginkortasunean zein faktorek eduki dezakete eragina?*

  5. eta 6. kapituluetan egindako esperimentuetan azterketa sakonak egin ditugu ahaidetasunean oinarritutako hedapen-teknika honen eraginkortasunean hainbat faktoreren eragina ikusteko. Besteak beste, honako eragile hauek aztertu ditugu: hainbat pa-

rametro (hedatutako kontzeptu edo termino kopurua, jatorrizko kontsultaren pisua, hedapeneko indizearen pisua) eta hauen optimizazioa, dokumentuen luzera, galderen zailtasuna eta datumultzoen nolakotasuna. Hauetako batzuentzat balio edo baliotarterik eraginkorrenak zein izan daitezkeen zehaztu daitekeela, baina beste ezaugarri batzuk datu-multzoarekiko edo erabilitako berreskurapen-ereduarekiko oso aldakorrak direla ikusi dugu. Gehiegi ez luzatzearren, hemen ez ditugu azalpen guzti horiek banan-banan zerrendatuko, 5.6 eta 6.6 ataletan dagoeneko zehaztu baititugu.

- 2.5 - *Hedapen-teknika hau egokia al da kontsulten eta dokumentuen itzulpena egiteko hizkuntza arteko berreskurapenean?*

  Hedapen-teknika honek WordNeten oinarritutako grafo-algoritmo bat erabiltzen du, zeinak *synset*ak itzultzen dituen. Ingeleseko esperimentuetan *synset* horiei dagozkien ingeleseko *variant*ak hartzen baditugu ere, beste edozein hizkuntzatako *variant*ak har genitzakeen. Hortaz, hedapen-teknika honek kontsulta eta dokumentuak itzultzeko balio dezake. Itzulpen-metodo hau erabili dugu Robust-WSD 2009 atazako hizkuntza arteko atazarako (gaztelania-ingelesa) prestatutako esperimentuetan.

7.2 Ekarpenak

Gure ekarpenik nagusiena ikerketa-galdera nagusiari erantzun positiboa ematea izan da: **semantika lexikalak hobetzen du IB sistemaren eraginkortasuna**. Kontsulta- eta dokumentu-hedapenerako HADa eta ahaidetasun semantikoa erabili ditugu, IB sistemaren eraginkortasunean positiboki eraginez. Jarraian ekarpenak zehaztasun handiagoz ikusiko ditugu, horietako bakoitza zein kapitulutan lortu den zehaztuz.

- **HAD informazioarekin aberastutako gai- eta dokumetu-bildumez baliatuz, kontsulta- eta dokumentu-hedapena egin ditugu** (4. kapitulua).

  Hedapena egiterakoan, kontsulta edo dokumentuko hitz bakoitzari bere sinonimoak gehitu dizkiogu ingelesezko eta gaztelaniazko WordNetez baliatuz. Kanpoko baliabide hau bakarrik erabiliz hedapenak egin,

eta parametroak doitu gabe, emaitza onak lortu ditugu. Gainera, hizkuntza arteko berreskurapenaren arloan, teknika honek kontsulten eta dokumentuen itzulpena egiteko balio duela ikusi dugu.

- **Ahaidetasun semantikoan oinarritutako teknika baten bidez, kontsulta eta dokumentuen hedapena egin dugu** (5. eta 6. kapituluak).

  Hedapen-teknika berritzaile bat proposatu dugu, zeinak WordNeten oinarritutako grafo-algoritmo baten bidez testu osoarekin erlazionatutako kontzeptuak —eta ondoren hitzak— lortzen dituen. Horrela, testuan bertan aipatzen ez diren, baina zerikusia duten kontzeptuak lortzen ditugu. Bi motako berreskurapen-ereduetan (eredu probabilistiko klasiko eta hizkuntza-ereduetan oinarritutako ereduan) txertatu dugu hedapen-teknika hau. Oro har, emaitza positiboak lortu ditugu, baita *query likelihood* eta PRF metodoekin alderatuz gero ere. Are gehiago, hizkuntza arteko berreskurapenaren arloan, teknika honek kontsulten eta dokumentuen itzulpena egiteko balio duela ikusi dugu.

- **Ahaidetasun semantikoan oinarritutako hedapen-teknikek izaera desberdineko datu-multzoekiko duten sendotasunaren azterketa egin dugu** (5. eta 6. kapituluak).

  Mota desberdineko datu-multzoetan eraginkortasuna nolakoa den aztertu nahi izan dugu. Horretarako, domeinu, kontsulta-tipologia eta dokumentu luzera desberdinetako hiru datu-multzo erabili ditugu gure esperimentuetan: (i) Robust, ad hoc atazetan erabili ohi den datu-multzo tipikoa, egunkarietako berriak dituena; (ii) Yahoo!, edozein gairen inguruan Interneteko erabiltzaileek adierazitako galdera eta erantzunak biltzen dituen datu-multzoa; eta (iii) ResPubliQA, pasarteen berreskurapenerako prestatutako Europar Batasuneko dokumentuez osatutako datu-multzoa. Esperimentuek erakusten dute gure hedapen-teknikak sendoak direla eta errendimendua ona dela hiru datu-multzoekin.

- **Ahaidetasun semantikoan oinarritutako hedapen-teknikek parametro-ezarpen desberdinekiko duten sendotasunaren azterketa gauzatu dugu** (5. eta 6. kapituluak).

  Parametroen optimizazioak eragin handia dauka IBko oinarri-lerroko sistemetan eta PRF metodoan; baita gure hedapen-metodoetan ere. Egoera erreal gehienetan datu-multzo berri batekin lanean hasi behar

dugunean askotan ez da egoten datu-multzo horren gainean parametroen doitzea egiteko entrenamendurako datu-multzorik. Gure hedapen-metodoak parametro-ezarpen ez-optimoekiko ere sendoak direla frogatu dugu.

- **Dokumentuen luzera eta ahaidetasunean oinarritutako hedapen-tekniken eraginkortasunaren artean loturarik ba ote dagoen aztertu dugu** (5. kapitulua).

  Dokumentuen batez besteko luzera desberdinetako sasibilduma batzuekin esperimentuak eginez, zenbat eta dokumentu motzagoak izan, ahaidetasun semantikoan oinarritutako hedapen-teknika orduan eta eraginkorragoa dela ikusi dugu, salbuespenak salbuespen.

- **Galderaren zailtasuna eta ahaidetasunean oinarritutako hedapen-tekniken eraginkortasunaren artean loturarik ba ote dagoen aztertu dugu** (6. kapitulua).

  Egindako azterketan PRF metodoa galdera errazentzat hobea eta guk proposatutako hedapen-teknikak galdera zailenetan eraginkorragoak direla ikusi dugu, bai eta konbinaziorako aukerak egon daitezkeela ere.

- **CLEF ebaluazio-kanpainaren baitako Robust-WSD (2008 eta 2009ko edizioetan) eta ResPubliQA atazetan (2009 eta 2010eko edizioetan) parte hartu dugu** (4. eta 5. kapituluak).

  Ataza horietako sailkapen orokorrean postu onak lortu genituenez, gure sistemak artearen egoeran kokatzen direla ikusi dugu. Modu berean, parte-hartze horrek beste parte-hartzaileen sistemekin konparatzeko aukera eman digu.

7.3 Etorkizuneko lanak

Batetik, tesi-lan honetan guztiz amaitu gabe gelditu diren ikerketa-lerroak, eta, bestetik, lan honi jarraipena emateko dauden ikerketa-lerroetako batzuk zerrendatuko ditugu jarraian:

- **Ahaidetasun semantikoan oinarritutako hedapen-eredua PRF-arekin konbinatu**.

Egindako analisiek ahaidetasun semantikoan oinarritutako hedapen-ereduak eta PRFa osagarriak direla erakutsi digute. Hori ikusita bi ereduak konbinatuz proba batzuk egin genituen, emaitza itxaropentsuak lortuz (6.5.5 atala). Ildo honi jarraituz, konbinaketa hori gehiago landu nahiko genuke.

- **Ahaidetasun semantikoaren bidez lortzen ditugun kontzeptuak aztertu, eta teknika hobetu**.

  Ahaidetasun semantikoa lortzeko grafo-algoritmoa parametro lehenetsiekin erabili dugu, ezarpen horiekin lortu baitzituzten hitzen antzekotasunerako emaitzarik onenak (Agirre *et al.*, 2009c). Algoritmo horren bidez lortzen ditugun kontzeptuen azaleko ebaluazioa besterik ez dugu egin. Hau gehiago aztertzea eta, beharrezkoa ikusten bada, ahaidetasun semantikorako teknika hau hobetzea nahiko genuke.

- **Ahaidetasun semantikoan oinarritutako hedapen-eredua berreskurapen-ereduan beste modu batera txertatu**.

  Proposatu dugun dokumentu-hedapenaren ereduan prozesu sinple bat jarraitzen dugu: hedapenetik lortutako hitzak bigarren indize batean sartzen ditugu, bi indizeei pisu desberdinak esleitzeko aukera emanez. Hedapenetik lortutako informazio guztia IB sisteman modu landuago batean txertatuz emaitza hobeak lortzen ote diren aztertu nahiko genuke. Horretarako, bi aukera aurreikusten ditugu. Alde batetik, BM25F berreskurapen-eredu probabilistikoarekin (Robertson *et al.*, 2004) esperimentatu ahalko genuke. Bestetik, Mei *et al.*-en (2008) eta Huang *et al.*-en (2009) lanetan oinarrituz, hizkuntza-ereduen leuntzea egiteko erabil genezake hedapeneko informazio hori.

- **Datu-multzo handiagoetarako eskalagarritasuna aztertu**.

  Gure esperimentuetan erabili dugun datu-multzo handienak milioi bat inguru dokumentu ditu. Gure ustez kopuru hauek egokiak dira egin nahi genituen ebaluazioetarako. Baina, bilatzailerik erabilienak web-bilatzaileak direnez, eta webaren tamaina geroz eta handiagoa denez, gero eta ohikoagoa da IB sistemak datu-multzo oso handietan ebaluatzea; ikusi besterik ez dago TREC ebaluazio-kanpainako azken edizioetako Web Track atazan erabili duten datu-multzo handienaren tamaina: bilioi bat web-orri[1]. Guk proposatutako hedapen-teknika horietan kos-

[1] http://plg.uwaterloo.ca/~trecweb/2011.html

tu handieneko eragiketa —denbora zein espazioari dagokionez— ahaidetasun semantikorako erabiltzen dugun grafo-algoritmoa da. Datumultzo handiagoekin esperimentuak egin nahi izanez gero, algoritmo honen eskalagarritasuna aztertu beharko genuke.

- **WordNet beharrean, beste ezagutza-iturriren bat erabili**.

  Proposatu dugun ahaidetasunean oinarritutako hedapen-teknika hori grafo-algoritmo eta ezagutza-base batean oinarritzen da. Gure esperimentuetarako aukeratu dugun ezagutza-basea WordNet izan da, grafo-algoritmo hori eta WordNet erabiliz beste esperimentu batzuetan emaitza onak lortu zirelako (Agirre *et al.*, 2009c; Agirre eta Soroa, 2009). WordNet izen arruntei eta aditzei dagokienez aberatsa bada ere, izen berezidun entitateen kopurua urria du. Kontuan izanik kontsultetan oso ohikoa dela horrelako entitateen bidez edo hauei buruz galdetzea, alegia, entitate hauek garrantzi handia izan dezaketela informazioaren berreskurapenean, horietako gehiago dituen ezagutza-iturriren bat erabiltzea ideia ona izan daitekeela iruditzen zaigu. Aukera bat WordNeten ordez edo WordNetekin batera Wikipedia erabiltzea izan daiteke. Wikipedian artikulu asko daude eta artikuluak beraien artean hiperesteka bidez erlazionatuta daude. Hori dela eta, bideragarria da Wikipedia grafo baten bidez errepresentatzea. Gainera, Wikipedia hitzen edo pasarteen arteko ahaidetasun semantikoa lortzeko erabili izan da emaitza onak lortuz (Milne eta Witten, 2008; Gabrilovich eta Markovitch, 2009).

- **Beste hizkuntza batzuekin probatu**.

  Sarreran esan dugun moduan, gure esperimentu gehienak ingelesarekin egin ditugu. Gaztelaniarekin ere egin ditugu esperimentu batzuk hizkuntza arteko informazioaren berreskurapeneko atazak direla eta. Hizkuntza arteko esperimentu hauetan emaitza itxaropentsuak lortu baditugu ere, itzulpen-teknika gehiago landu beharra dagoela uste dugu. Bestetik, euskarazko datu-multzoren batekin ebaluazioren bat egitea gustatuko litzaiguke. Interesgarria izango litzateke zientzia eta teknologiaren domeinura mugatzea esperimentuak eta WNTERM ontologia espezializatua (Pociello *et al.*, 2008) erabiltzea ezagutza-base moduan.

Bibliografia

Agirre E. *Kontzeptuen arteko erlazio-izaeraren formalizazioa ontologiak erabiliaz: dentsitate kontzeptuala.* Doktoretza-tesia, Informatika Fakultatea, UPV-EHU, 1999.

Agirre E., Aldezabal I., eta Pociello E. Euskararako ezagutza-base lexiko-semantikoaren eredu-hautaketa eta garapena: EuskalWordNet. *GOGOA aldizkaria ISNN 1577-9424*, 237–266, 2006.

Agirre E., Ansa O., Arregi X., Lopez de Lacalle M., Otegi A., eta Saralegi X. Document expansion for cross-lingual passage retrieval. *Proceedings of CLEF 2010 Workshop on Multiple Language Question Answering (MLQA'10)*, 2010a. ISBN 978-88-904810-0-0.

Agirre E., Ansa O., Arregi X., Lopez de Lacalle M., Otegi A., Saralegi X., eta Zaragoza H. Elhuyar-IXA: semantic relatedness and cross-lingual passage retrieval. *Multilingual Information Access Evaluation I. Text Retrieval Experiments, CLEF 2009*, 6241 lib. of *Lecture Notes in Computer Science*, 273–280. Springer, 2010b. ISBN 978-3-642-15753-0.

Agirre E., Cuadros M., Rigau G., eta Soroa A. Exploring knowledge bases for similarity. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010c. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Agirre E., Di Nunzio G.M., Ferro N., Mandl T., eta Peters C. CLEF 2008: ad hoc track overview. *Evaluating Systems for Multilingual and Multimodal Information Access, CLEF 2008*, 5706 lib. of *Lecture Notes in Computer Science*, 15–37. Springer, 2009a. ISBN 978-3-642-04446-5.

Agirre E., Di Nunzio G.M., Mandl T., eta Otegi A. CLEF 2009 Ad Hoc Track Overview: Robust-WSD Task. *Multilingual Information Access Evaluation I. Text Retrieval Experiments, CLEF 2009*, 6241 lib. of *Lecture Notes in Computer Science*, 36–49. Springer, 2010d. ISBN 978-3-642-15753-0.

Agirre E. eta Edmonds P., editors. *Word sense disambiguation: algorithms and applications*, Text, Speech and Language Technology Series, 33 lib. Springer, 2006.

Agirre E. eta Lopez de Lacalle O. UBC-ALM: combining k-NN with SVD for WSD. *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, 342–345, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

Agirre E., Lopez De Lacalle O., Magnini B., Otegi A., Rigau G., eta Vossen P. *Advances in multilingual and multimodal information retrieval*, chapter SemEval-2007 Task 01: Evaluating WSD on Cross-Language Information Retrieval, 908–917. Springer-Verlag, 2008. ISBN 978-3-540-85759-4.

Agirre E., Otegi A., eta Rigau G. IXA at CLEF 2008 Robust-WSD task: using word sense disambiguation for (cross lingual) information retrieval. *Evaluating Systems for Multilingual and Multimodal Information Access, CLEF 2008*, 5706 lib. of *Lecture Notes in Computer Science*, 118–125. Springer, 2009b. ISBN 978-3-642-04446-5.

Agirre E., Otegi A., eta Zaragoza H. Using semantic relatedness and word sense disambiguation for (CL)IR. *Multilingual Information Access Evaluation I. Text Retrieval Experiments, CLEF 2009*, 6241 lib. of *Lecture Notes in Computer Science*, 166–173. Springer, 2010e. ISBN 978-3-642-15753-0.

Agirre E. eta Soroa A. Personalizing PageRank for word sense disambiguation. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, 33–41. Association for Computational Linguistics, 2009.

Agirre E., Soroa A., Alfonseca E., Hall K., Kravalova J., eta Paşca M. A study on similarity and relatedness using distributional and WordNet-based approaches. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, 19–27, Stroudsburg, PA, USA, 2009c. Association for Computational Linguistics. ISBN 978-1-932432-41-1.

Altingövde I.S., Demir E., Can F., eta Ulusoy O. Incremental cluster-based retrieval using compressed cluster-skipping inverted files. *ACM Trans. Inf. Syst.*, 26:15:1–15:36, 2008. ISSN 1046-8188.

Amati G. eta van Rijsbergen C.J. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.

Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., eta Ives Z. DBpedia: a nucleus for a web of open data. *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ISWC'07/ASWC'07, 722–735. Springer-Verlag, 2007. ISBN 3-540-76297-3, 978-3-540-76297-3.

Baeza-Yates R. eta Ribeiro-Neto B. *Modern information retrieval - the concepts and technology behind search, Second edition.* Pearson Education Ltd., Harlow, England, 2011. ISBN 978-0-321-41691-9. http://www.mir2ed.org/.

Bates M.J. Subject access in online catalog: a design model. *Journal of The American Society for Information Science*, 357–376, 1986.

Blair D.C. eta Maron M.E. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *CACM*, 28(3):289–299, 1985.

Boldi P. eta Vigna S. MG4J at TREC 2005. *The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings*, number SP 500-266 in Special Publications. NIST, 2005. http://mg4j.dsi.unimi.it/.

Brin S. eta Page L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

Buckley C. eta Sanderson M. Relevance Feedback Track Overview: TREC 2008. *Proceedings of The Seventeenth Text REtrieval Conference, TREC*

2008, Special Publication 500-277 lib. National Institute of Standards and Technology (NIST), 2008.

Budanitsky A. eta Hirst G. Evaluating WordNet-based measures of lexical semantic relatedness. *Computacional Linguistics*, 32:13–47, 2006. ISSN 0891-2017.

Bush V. As we may think. *The Atlantic Monthly*, 176(1):101–108, 1945.

Can F., Altingövde I.S., eta Demir E. Efficiency and effectiveness of query processing in cluster-based retrieval. *Information Systems*, 29(8):697–717, 2004. ISSN 0306-4379.

Can F. eta Ozkarahan E.A. Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases. *ACM Trans. Database Syst.*, 15(4):483–517, 1990.

Chan Y.S., Ng H.T., eta Zhong Z. NUS-PT: exploiting parallel texts for word sense disambiguation in the English all-words tasks. *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, 253–256, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

Cleverdon C.W. The significance of the Cranfield tests on index languages. *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '91, 3–12, New York, NY, USA, 1991. ACM. ISBN 0-89791-448-1.

Cleverdon C.W., Mills J., eta Keen M. Factors determining the performance of indexing systems. *ASLIB Cranfield project, Cranfield*, 1966.

Croft W.B. A file organization for cluster-based retrieval. *Proceedings of the 1st annual international ACM SIGIR conference on Information storage and retrieval*, SIGIR '78, 65–82. ACM, 1978.

Croft W.B., Metzler D., eta Strohman T. *Search engines: information retrieval in practice*. Addison-Wesley Publishing Company, USA, 1st edition, 2009. ISBN 0136072240, 9780136072249.

Fellbaum C. *WordNet: an electronic lexical database and some of its applications*. MIT Press, Cambridge, Mass, 1998.

Forner P., Peñas A., Agirre E., Alegria I., Forăscu C., Moreau N., Osenova P., Prokopidis P., Rocha P., Sacaleanu B., Sutcliffe R., eta Sang E. Overview of the CLEF 2008 multilingual question answering track. *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, 262–295. Springer-Verlag, 2009. ISBN 3-642-04446-8, 978-3-642-04446-5.

Fox C. *Information retrieval*, chapter Lexical analysis and stoplists, 102–130. Prentice-Hall, Inc., 1992. ISBN 0-13-463837-9.

Frakes W.B. *Information retrieval: data structures and algorithm*, chapter Stemming algorithms, 131–160. Prentice-Hall, Inc., 1992. ISBN 0-13-463837-9.

Furnas G.W., Landauer T.K., Gomez L.M., eta Dumais S.T. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

Gabrilovich E. eta Markovitch S. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34(1):443–498, 2009. ISSN 1076-9757.

Giunchiglia F., Kharkevich U., eta Zaihrayeu I. Concept Search. *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC 2009, 429–444. Springer-Verlag, 2009. ISBN 978-3-642-02120-6.

Gonzalo J., Verdejo F., Chugur I., eta Cigarran J. Indexing with WordNet synsets can improve text retrieval. *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, 38–44, 1998.

Greiff W.R., Croft W.B., eta Turtle H. PIC matrices: a computationally tractable class of probabilistic query operators. *ACM Trans. Inf. Syst.*, 17 (4):367–405, 1999.

Harman D. How effective is suffixing? *JASIS*, 42:7–15, 1991.

Haveliwala T.H. Topic-sensitive PageRank. *Proceedings of WWW '02*, 517–526, 2002. ISBN 1-58113-449-5.

Hiemstra D. *Information retrieval: searching in the 21st century*, chapter Information Retrieval Models, 1–19. John Wiley & Sons, Ltd, 2009. ISBN 9780470033647.

Hollink V., Kamps J., Monz C., eta de Rijke M. Monolingual document retrieval for European languages. *Information Retrieval*, 7(1):33–52, 2004.

Hsu M., Tsai M., eta Chen H. Combining WordNet and ConceptNet for automatic query expansion: a learning approach. *Proceedings of the 4th Asia information retrieval conference on Information retrieval technology*, AIRS'08, 213–224. Springer-Verlag, 2008. ISBN 3-540-68633-9, 978-3-540-68633-0.

Huang Y., Sun L., eta Nie J. Smoothing document language model with local word graph. *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, 1943–1946. ACM, 2009. ISBN 978-1-60558-512-3.

Hughes T. eta Ramage D. Lexical semantic relatedness with random graph walks. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 581–589, 2007.

Hull D. Using statistical testing in the evaluation of retrieval experiments. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, 329–338, New York, NY, USA, 1993. ACM. ISBN 0-89791-605-0.

Hull D. Stemming algorithms – A case study for detailed evaluation. *JASIS*, 47(1):70–84, 1996.

Jardine N. eta van Rijsbergen C.J. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.

Jones K.S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

Jones K.S., Walker S., eta Robertson S.E. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36:779–808, 2000. ISSN 0306-4573.

Kasneci G., Ramanath M., Suchanek F., eta Weikum F. The YAGO-NAGA approach to knowledge discovery. *SIGMOD Rec.*, 37(4):41–47, 2009. ISSN 0163-5808.

Kekäläinen J. eta K. Järvelin K. The impact of query structure and query expansion on retrieval performance. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, 130–137. ACM, 1998. ISBN 1-58113-015-5.

Kent A., Berry M.M., Luehrs Jr F.U., eta Perry J.W. Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American Documentation*, 6(2):93–101, 1955.

Kim S., Seo H., eta Rim H. Information retrieval using word senses: root sense tagging approach. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, 258–265, New York, NY, USA, 2004. ACM. ISBN 1-58113-881-4.

Krovetz R. eta Croft W.B. Lexical ambiguity and information retrieval. *ACM Trans. Inf. Syst.*, 10(2):115–141, 1992.

Kurland O. eta Lee L. Corpus structure, language models, and ad hoc information retrieval. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, 194–201, New York, NY, USA, 2004. ACM. ISBN 1-58113-881-4.

Lavrenko V. eta Croft W.B. 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, NY, USA, 2001. ACM. ISBN 1-58113-331-6.

Leturia I., Gurrutxaga A., Areta N., Alegria I., eta Ezeiza A. EusBila, a search service designed for the agglutinative nature of Basque. *SIGIR2007-iNEWS'07 workshop*, 2007.

Leturia I., Gurrutxaga A., Areta N., eta Pociello E. Analysis and performance of morphological query expansion and language-filtering words on Basque web searching. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), 2008. ISBN 2-9517408-4-0.

Liu H. eta Singh P. ConceptNet – A practical commonsense reasoning toolkit. *BT Technology Journal*, 22(4):211–226, 2004. ISSN 1358-3948.

Liu S., Liu F., Yu C., eta Meng W. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, 266–272, New York, NY, USA, 2004. ACM. ISBN 1-58113-881-4.

Liu S., Yu C., eta Meng W. Word sense disambiguation in queries. *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, 525–532, New York, NY, USA, 2005. ACM. ISBN 1-59593-140-6.

Liu X. eta Croft W.B. Cluster-based retrieval using language models. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, 186–193, New York, NY, USA, 2004. ACM. ISBN 1-58113-881-4.

Lopez de Lacalle O. *Domain-specific word sense disambiguation*. Doktoretza-tesia, Lengoiaia eta Sistema Informatikoak Saila, UPV/EHU, 2009.

Lopez de Lacalle O. eta Agirre E. Hitzen adiera-desanbiguazioa. *EKAIA 23*, 127–156, 2010.

Luhn H.P. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4): 309–317, 1957. ISSN 0018-8646.

Manning C.D., Raghavan P., eta Schütze H. *An introduction to information retrieval*. Cambridge University Press, UK, 2009.

Maron M.E. eta Kuhns J.L. On relevance, probabilistic indexing and information retrieval. *J. ACM*, 7(3):216–244, 1960. ISSN 0004-5411.

Martinez D. *Supervised word sense disambiguation: facing current challenges*. Doktoretza-tesia, Informatika Fakultatea, UPV/EHU, 2004.

Mei Q., Zhang D., eta Zhai C. A general optimization framework for smoothing language models on graph structures. *Proceedings of the 31st*

annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '08, 611–618. ACM, 2008. ISBN 978-1-60558-164-4.

Meij E. *Combining concepts and language models for information access.* Doktoretza-tesia, Informatics Institute, University of Amsterdam, 2010.

Metzler D. eta Croft W.B. Combining the language model and inference network approaches to retrieval. *Inf. Process. Manage.*, 40:735–750, 2004. ISSN 0306-4573.

Mihalcea R. eta Moldovan D. Semantic indexing using WordNet senses. *Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval*, 11 lib. of *RANLPIR '00*, 35–45. ACL, 2000.

Miller G.A., Leacock C., Tengi R., eta Bunker R.T. A semantic concordance. *Proceedings of the workshop on Human Language Technology*, HLT '93, 303–308, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. ISBN 1-55860-324-7.

Milne D. eta Witten I.H. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. *Proceedings of the first AAAI Workshop on Wikipedia and Artifical Intellegence (WIKIAIâĂŹ08)*, 2008.

Mitra M., Singhal A., eta Buckley C. Improving automatic query expansion. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, 206–214, New York, NY, USA, 1998. ACM. ISBN 1-58113-015-5.

Monarch I. eta Carbonell J. CoalSORT: a knowledge-based interface. *IEEE Expert*, 2:39–53, 1987.

Mooers C.N. Information retrieval viewed as temporal signaling. *Proceedings of the International Congress of Mathematicians*, 1950.

Muller C. eta Gurevych I. 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, 2009. Association for Computational Linguistics.

Navigli R. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69, 2009. ISSN 0360-0300.

O'Connor J. Some remarks on mechanized indexing and some small-scale empirical results. *Machine Indexing: Progress and Problems*, 262–279. The American University, 1961.

Peñas A., Forner P., Rodrigo A., Sutcliffe R., Forăscu C., eta Mota C. Overview of ResPubliQA 2010: question answering evaluation over european legislation. *Proceedings of CLEF 2010 Workshop on Multiple Language Question Answering (MLQA'10)*, 2010. ISBN 978-88-904810-0-0.

Peñas A., Forner P., Sutcliffe R., Rodrigo A., Forăscu C., Alegria I., Giampiccolo D., Moreau N., eta Osenova P. Overview of ResPubliQA 2009: question answering evaluation over European legislation. *Proceedings of the 10th cross-language evaluation forum conference on Multilingual information access evaluation: text retrieval experiments*, CLEF'09, 174–196, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 3-642-15753-X, 978-3-642-15753-0.

Pérez-Agüera J. eta Zaragoza H. Query clauses and term independence. *Evaluating Systems for Multilingual and Multimodal Information Access, CLEF 2008*, 5706 lib. of *Lecture Notes in Computer Science*, 138–145. Springer, 2009. ISBN 978-3-642-04446-5.

Pociello E. *Euskararen ezagutza-base lexikala: Euskal WordNet*. Doktoretza-tesia, Euskal Filologia Saila, UPV/EHU, 2008.

Pociello E., Agirre E., eta Aldezabal I. Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 1–22, 2010. ISSN 1574-020X.

Pociello E., Gurrutxaga A., Agirre E., Aldezabal I., eta Rigau G. WNTERM: enriching the MCR with a terminological dictionary. *Proceedings of the 6th International Conference on Language Resources and Evaluations (LREC)*, 2008.

Ponte J.M. eta Croft W.B. 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, NY, USA, 1998. ACM. ISBN 1-58113-015-5.

Porter M.F. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

Pradhan S.S., Loper E., Dligach D., eta Palmer M. SemEval-2007 task 17: English lexical sample, SRL and all words. *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, 87–92, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

Rada R., Mili H., Bicknell E., eta Blettner M. Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions*, 19:17–30, 1989.

Resnik P. WSD in NLP Applications. *Word sense disambiguation: algorithms and applications*, 33 lib. of *Text, Speech and Language Technology*, 299–338. Springer, Dordrecht, The Netherlands, 2006.

Richardson R. eta Smeaton A.F. Using WordNet in a knowledge-based approach to information retrieval. Barne-txostena CA-0395, Dublin City University, Dublin, Ireland, 1995.

Robertson S. Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520, 2004.

Robertson S. On GMAP: and other transformations. *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, 78–83, New York, NY, USA, 2006. ACM. ISBN 1-59593-433-2.

Robertson S. eta Zaragoza H. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. ISSN 1554-0669.

Robertson S., Zaragoza H., eta Taylor M. 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, NY, USA, 2004. ACM. ISBN 1-58113-874-1.

Robertson S.E. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

Robertson S.E. eta Jones K.S. Relevance weighting of search terms. *JASIS*, 27(3):129–146, 1976. ISSN 1097-4571.

Robertson S.E. eta Walker S. 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, 1994. Springer-Verlag New York, Inc. ISBN 0-387-19889-X.

Rocchio J.J. Relevance feedback in information retrieval. *The Smart retrieval system - experiments in automatic document processing*, 313–323. Englewood Cliffs, NJ: Prentice-Hall, 1971.

Ruthven I. eta Lalmas M. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, 2003. ISSN 0269-8889.

Sag I.A., Baldwin T., Bond F., Copestake A.A., eta Flickinger D. Multiword expressions: a pain in the neck for NLP. *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, 1–15. Springer-Verlag, 2002. ISBN 3-540-43219-1.

Salton G. Cluster search strategies and the optimization of retrieval effectiveness. *The SMART retrieval system – Experiments in automatic document processing*, 223–242. Prentice Hall, 1971a.

Salton G. *The SMART retrieval system – Experiments in automatic document processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971b.

Salton G. *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Addison Wesley, Reading, MA, 1989.

Salton G. eta Buckley C. On the use of spreading activation methods in automatic information. *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '88, 147–160. ACM, 1988. ISBN 2-7061-0309-4.

Sanderson M. Word sense disambiguation and information retrieval. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, 142–151, New York, NY, USA, 1994. Springer-Verlag New York, Inc. ISBN 0-387-19889-X.

Sanderson M. *Word sense disambiguation and information retrieval*. Doktoretza-tesia, Department of Computing Science at the University of Glasgow, 1997.

Sanderson M. Retrieving with good sense. *Information Retrieval*, 2(1):49–69, 2000.

Sanderson M. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4:247–375, 2010.

Saralegi X. eta Lopez de Lacalle M. Dictionary and monolingual corpus-based query translation for Basque-English CLIR. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*. European Language Resources Association, 2010. ISBN 2-9517408-6-7.

Schutze H. eta Pedersen J.O. Information retrieval based on word senses. *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, 161–175, 1995.

Singhal A. eta Pereira F. Document expansion for speech retrieval. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, 34–41, New York, NY, USA, 1999. ACM. ISBN 1-58113-096-1.

Singitham P.K.C., Mahabhashyam M.S., eta Raghavan P. Efficiency-quality tradeoffs for vector score aggregation. *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*, VLDB '04, 624–635. VLDB Endowment, 2004. ISBN 0-12-088469-0.

Smeaton A.F. eta Quigley I. Experiments on using semantic distances between words in image caption retrieval. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, 174–180. ACM, 1996. ISBN 0-89791-792-8.

Smucker M.D., Allan J., eta Carterette B. A comparison of statistical significance tests for information retrieval evaluation. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, 623–632, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9.

Stokoe C., Oakes M.P., eta Tait J. Word sense disambiguation in information retrieval revisited. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, 159–166. ACM, 2003. ISBN 1-58113-646-3.

Strohman T., Metzler D., Turtle H., eta Croft W.B. Indri: a language-model based search engine for complex queries. Barne-txostena, Proceedings of the International Conference on Intelligent Analysis, 2005.

Surdeanu M., Ciaramita M., eta Zaragoza H. Learning to rank answers on large online QA collections. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, 719–727. The Association for Computer Linguistics, 2008. ISBN 978-1-932432-04-6.

Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network. *Proceedings of the second international conference on Information and knowledge management*, CIKM '93, 67–74. ACM, 1993. ISBN 0-89791-626-3.

Swanson D.R. Historical note: information retrieval and the future of an illusion. *JASIS*, 39(2):92–98, 1988.

Tao T., Wang X., Mei Q., eta Zhai C. Language model information retrieval with document expansion. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, 407–414. Association for Computational Linguistics, 2006.

Turtle H. eta Croft W.B. Evaluation of an inference network-based retrieval model. *ACM Trans. Inf. Syst*, 9:187–222, 1991. ISSN 1046-8188.

van Rijsbergen C.J. *Information retrieval*. Butterworths, London, 2nd edition, 1979.

Varelas G., Voutsakis E., Raftopoulou P., Petrakis E.G.M., eta Milios E.E. Semantic similarity methods in wordNet and their application to information retrieval on the web. *Proceedings of the 7th annual ACM international workshop on Web information and data management*, WIDM '05, 10–16. ACM, 2005. ISBN 1-59593-194-5.

Voorhees E.M. The cluster hypothesis revisited. *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '85, 188–196. ACM, 1985. ISBN 0-89791-159-8.

Voorhees E.M. Using WordNet to disambiguate word senses for text retrieval. *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, 171–180, New York, NY, USA, 1993. ACM. ISBN 0-89791-605-0.

Voorhees E.M. Query expansion using lexical-semantic relations. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, 61–69, New York, NY, USA, 1994. Springer-Verlag New York, Inc. ISBN 0-387-19889-X.

Voorhees E.M. Natural language processing and information retrieval. *Information Extraction: Towards Scalable, Adaptable Systems*, 32–48, London, UK, 1999. Springer-Verlag. ISBN 3-540-66625-7.

Voorhees E.M. The TREC robust retrieval track. *SIGIR Forum*, 39:11–20, 2005. ISSN 0163-5840.

Vossen P., editor. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA, 1998. ISBN 0-7923-5295-5.

Wallis P. Information retrieval based on paraphrase. *Proceedings of PACLING Conference*, 1993.

Weiss S.F. Learning to disambiguate. *Information Storage and Retrieval*, 9 (1):33–41, 1973.

Xu J. eta Croft W.B. Query expansion using local and global document analysis. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, 4–11, New York, NY, USA, 1996. ACM. ISBN 0-89791-792-8.

Xu J. eta Croft W.B. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18:79–112, 2000. ISSN 1046-8188.

Zernik U. Train1 vs. Train2: tagging word senses in corpus. *Proceedings of RIAO 91, Intelligent Text and Image Handling*, 567–585, 1991.

Zhai C. Statistical language models for information retrieval a critical review. *Foundations and Trends in Information Retrieval*, 2(3):137–213, 2008. ISSN 1554-0669.

Zhai C. eta Lafferty J. A study of smoothing methods for language models applied to ad hoc information retrieval. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, 334–342, New York, NY, USA, 2001a. ACM. ISBN 1-58113-331-6.

Zhai C. eta Lafferty J. Model-based feedback in the language modeling approach to information retrieval. *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, 403–410, New York, NY, USA, 2001b. ACM. ISBN 1-58113-436-3.

Glosategia

ad hoc ataza (*ad hoc task*)

IBko atazarik arruntena. Ataza honetan erabiltzaileak informazio-behar bat du eta behar hori asetzeko bilaketa bat egiten du dokumentu-bilduma baten gainean IB sistema bat erabiliaz. Sistema honek kontrolpean duen dokumentu-bilduma horretatik informazio-behar horrentzako adierazgarriak diren dokumentuak itzuliko dizkio erabiltzaileari.

adierazgarritasun epaiak (*relevance judgments, qrels*)

Zerrenda bat, non informazio-behar bakoitzerako zein dokumentu diren adierazgarriak (eta zeintzuk ez) zehazten den.

agerkidetza (*co-ocurrence*)

Dokumentu batean bi termino edo gehiago elkarren ondoan izateari deritzo, zoriz lortutakoa baino handiagoa den maiztasun batekin.

ahaidetasun semantiko (*semantic relatedness*)

Antzekotasun semantikoa baino orokorragoa den kontzeptua, sinonimia eta hiperonimia/hiponimia erlazioez gain, hitzen arteko meronimia (esku-hatz), antonimia (hotz-bero) eta asoziazioa (arkatza-papera) bezalako erlazioak ere kontuan hartzen dituena.

antzekotasun semantiko (*semantic similarity*)

Hitzen arteko sinonimia (auto-beribil) eta hiperonimia/hiponimia (taxi-auto) erlazioak bere baitan hartzen dituen kontzeptua.

AP (*average precision*)

IB sistema bati egindako kontsulta baten eraginkortasuna neurtzeko neurri bat. Kontsulta batentzat berreskuratutako dokumentu adierazgarri guztien posizioei dagokien doitasunaren batez bestekoa kalkulatzen du. Neurri honek dokumentu adierazgarri guztien posizioak hartzen ditu kontuan, baina eragin handia dute lehen postuetako dokumentu adierazgarriek.

bag of words

IBan erabiltzen den eredu bat. Eredu honetan dokumentuak hitzez betetako zaku bezala ikus daitezke; hau da, dokumentuan hitzek duten segida erabat galtzen da, eta, hortaz, ordena ez da kontuan hartzen bilaketak egiterakoan.

berreskurapen-eredu (*retrieval model*)

Erabiltzailearen kontsulta emanik, erabiltzailearentzat adierazgarria zer izango den aurreikusten eta azaltzen du. Oinarri matematiko batzuk erabiliz formalizatzen da, eta IB sistemek hori inplementatzen dute.

bilatzaile (*search engine*)

Konputagailuetan informazioa bilatzeko garatutako IB sistema bat. Erabiltzaileak kontsulta baten bidez sistemari zer bilatu nahi duen adierazten dio, eta sistemak kontsulta horren araberako elementuak itzuliko dizkio.

datu-multzo (*dataset*)

IB sistema baten ebaluazioa egiteko erabiltzen den datu-bilduma bat, hainbat galdera, dokumentu eta adierazgarritasun-epaiez osatua egongo dena.

doitasun (*precision*)

IB sistema baten eraginkortasuna neurtzeko oinarrizko neurrietako bat. Kontsulta batentzat berreskuratutako dokumentu-bildumatik adierazgarriak zenbat diren adierazten du.

dokumentu adierazgarri (*relevant document*)

Erabiltzailearen informazio-beharra asetuko duen dokumentua; alegia, kontsultan adierazitako informazioa duen dokumentua.

dokumentu-maiztasunaren alderantzikatua (*inverse document frequency*, idf)

IB arloan asko erabiltzen den neurri estatistiko bat, *idf* moduan adierazi ohi dena. Neurri honek termino bat bildumako zenbat dokumentutan agertzen den kontatu eta balio horren alderantzikatua adierazten du. Hortaz, *idf* altua izango du bilduma osoan agerpen gutxi dituen terminoak, eta, alderantziz, baxua izango du oso ohikoa den terminoak.

entrenamendurako bilduma (*training collection*)

Datu-multzo osoaren zati bat, parametroak doitzeko erabiliko dena.

eraginkortasun (*effectiveness*)

IB sistema batek informazio egokia bilatzeko duen gaitasuna.

eredu probabilistiko (*probabilistic model*)

IB sistema batek jarraitu dezakeen ereduetako bat. Eredu honen helburua Q kontsultaren arabera D dokumentuak $P(R = 1|Q, D)$ probabilitatearen arabera ordenatzea izango da. Printzipio honetan adierazgarritasuna (ingelesezko relevancetik R moduan adierazi ohi dena) aldagai bitartzat hartzen da, eta, beraz, $R = 1$ izango da D dokumentua adierazgarria denean Q kontsultarentzako, eta 0 bestela.

errendimendu (*efficiency*)

IB sistema baten bilaketen azkartasuna adierazten duen neurria.

erro-bilatzaile (*stemmer*)

Hitz bakoitzaren erroa zein den esaten duen tresna. Hori erregela finko batzuen bidez ohikoenak diren atzizkiak kenduz egiten du, ez du baliabide edo prozesu linguistikorik erabiltzen lematizatzaileak bezala.

esangura-test (*significance test*)

Bi sistemen ebaluazioen artean dauden diferentziak estatistikoki esanguratsuak diren edo ez esaten duen metodo edo froga (beti ere hutsegite probabilitate bat kontuan izanik).

estaldura (*recall*)

IB sistema baten eraginkortasuna neurtzeko oinarrizko neurrietako bat. Kontsulta batentzat berreskuratutako dokumentu adierazgarrien zatia zenbatekoa den adierazten du.

estatistikoki esanguratsu (*statistically significant*)

Estatistikan emaitza bat estatistikoki esanguratsua dela esaten da zoriz gertatzeko aukerarik egon ez denean.

exekuzio (*run*)

IB sistema batek kontsulta-sorta bat egikaritu ondoren, irteera moduan, kontsulta bakoitzerako itzulitako dokumentuak bilduko dituen fitxategia.

ezagutza-base lexikal (*lexical knowledge-base*)

Hitz eta adierei buruzko informazioa duen biltegi edo lexikoia. Lexikoi hauen ezaugarri garrantzitsuena herentzia izaten da, adierak klase/azpiklase hierarkien inguruan antolatzen baitira.

gai (*topic*)

Robust datu-multzoan informazio-beharra adierazteko erabiltzen den testua, beste bildumetako galderaren parekoa izango litzatekeena.

gako-hitz (*keyword*)

Bilaketa bat egiteko erabiltzen den kontsultako hitzetako bakoitza. Ohikoena informazio-beharra lengoaia naturalean adierazitako esaldi arrunt edo galdera bat izatea da. Kontsulta esaldi horretako gako-hitzak diren terminoak hartuz osatuko da.

GMAP (*geometric mean average precision*)

IB sistema baten eraginkortasuna neurtzeko neurri bat. AParen batez besteko geometrikoa da. Batez besteko doitasunen biderkadura egiten da, eta horrek galdera zailetan egindako hobekuntzen eragina nabarmentzen du emaitzan.

HAD

Ikus *hitzen adiera-desanbiguazio*.

hedapen (*expansion*)

Testu zati bati hitz berriak gehitzeko teknika, beti ere, hitz horiek testuko hitzekin nolabaiteko erlazio edo ahaidetasun semantikoren bat dutelarik. Bai kontsulten eta bai dokumentuen hedapena egin daiteke.

hiperonimia (*hypernymy*)

Kontzeptu orokorrenak kontzeptu zehatzagoekin lotzen dituen erlazio semantikoa. Adibidez, auto hitza taxi hitzarekiko hiperonimiako erlazioan dago.

hiponimia (*hyponymy*)

Kontzeptu zehatzenak kontzeptu orokorragoekin lotzen dituen erlazio semantikoa. Adibidez, taxi hitza auto hitzarekiko hiponimiako erlazioan dago.

hitzen adiera-desanbiguazio, HAD (*word sense disambiguation*, WSD)

Konputazio-metodoak erabiliz hitzen agerpenei adiera egokia esleitzen dien prozesua.

hizkuntza-eredu (*language model*)

Hitz-segidei probabilitate-banaketak esleitzen dizkien eredu probabilistiko bat. Sinpleena unigrametako hizkuntza-eredua da, non estimazioak egiterakoan hitz bakoitza besteekiko independentetzat hartzen den. Hizkuntzaren prozesamenduan asko erabiltzen da, eta baita IBan ere.

hizkuntza arteko ataza (*cross-lingual task*)

Hizkuntza batean dagoen kontsultaren bidez beste hizkuntza batean (edo bat baino gehiagotan) dauden dokumentuak berreskuratzen diren ataza.

hizkuntzaren prozesamendu, HP (*natural language processing*, NLP)

Hizkuntzaren tratamendu automatikoaren inguruko ikerrarloa.

HP

Ikus *hizkuntzaren prozesamendu*.

IB

Ikus *informazioaren berreskurapen*.

idf

Ikus *dokumentu-maiztasunaren alderantzikatua*.

indize (*index*)

Bilaketa azkarrak egitea ahalbideratuko duen datu-egitura bat. Datu-egitura honetan hitzak zein dokumentuetan agertu diren gordetzen da, eta batzuetan dokumentuko bakoitzean zein posizioetan dagoen ere gordeko da.

informazio-behar (*information need*)

Erabiltzaileak, gai baten inguruan gehiago jakin nahi duelako, bilaketa bat egiterakoan egiten duen adierazpena, askotan lengoaia naturalean adierazia.

informazioaren berreskurapen, IB (*information retrieval*, IR)

Erabiltzaile baten informazio-beharra asetuko duten dokumentuak bilatzeko prozesua. Prozesu honetan erabiltzaileak bere informazio-beharra lengoaia naturaleko adierazpen baten bidez edo termino solte batzuen bidez adieraziko dio IB sistema bati. IB sistemak informazio-behar horretan oinarritutako kontsulta bat sortuko du, eta hor eskatzen den informazioa eduki dezaketen dokumentuen berri emango dio erabiltzaileari.

kontsulta (*query*)

Erabiltzailearen informazio-beharra sistemari adierazteko erabiliko den hitz-segida. Ohikoena informazio-beharra lengoaia naturalean adierazitako esaldi arrunt edo galdera bat izatea da eta kontsulta esaldi horretako gako-hitzak diren terminoak hartuz osatuko da.

kontsulta egituratu (*structured query*)

Termino-segida soil bat izan beharrean, terminoak multzokatuta, terminoen arteko erlazioak edota pisu desberdineko terminoak izan ditzakeen kontsulta.

kontsulta-egiantza (*query likelihood*, QL)

IBrako hizkuntza-eredu probabilistiko bat. Eredu honetan bildumako D dokumentu bakoitzaren Θ_D hizkuntza-eredua sortzen da eta Θ_D estimazio horretan oinarrituta, D dokumentutik Q kontsulta sortu izanaren egiantzaren arabera ordenatzen dira dokumentuak berreskuratze-prozesuan.

lema (*lemma*)

Hitza atzizki flexiborik gabe, hiztegietan sarrera gisa dagoen hori.

lematizatzaile (*lemmatizer*)

Hitz bakoitzaren lema zein den esaten duen aplikazioa. Horretarako, erro-bilatzaileek ez bezala, tresna honek baliabide edo prozesu linguistikoak erabiltzen ditu.

leuntze (*smoothing*)

Besteak beste, dokumentuan agertzen ez diren terminoei zero probabilitatea esleitu beharrean, probabilitate-masa txiki bat esleitzeko teknika. Hitz gutxitan esanda, gertaera ezagunentzat estimatutako probabilitatea txikiagotu eta gertaera ezezagunentzat estimatutako probabilitatea handiagotzen du teknika honek.

MAP (*mean average precision*)

IB sistema baten eraginkortasuna neurtzeko neurririk erabilienetako bat. Exekuzio oso bat ebaluatzeko erabiltzen da, kontsulta bakoitzaren AParen batez bestekoa kalkulatuz.

MRR (*mean reciprocal rank*)

IB sistema baten eraginkortasuna neurtzeko neurri bat. *Reciprocal rank* deiturikoak lehen dokumentu adierazgarria zein posiziotan berreskuratzen duen hartzen du kontuan, hots, bigarren postuan berreskuratzen bada dokumentu adierazgarria, $1/2 = 0,5$ izango da bere balioa. Eta MRRak kontsulta guztien balio horren batez bestekoa kalkulatzen du.

oinarri-lerroko sistema (*baseline system*)

Lantzen ari den arazoaren soluzio sinple bat, oinarritzat hartu ohi dena emaitzen konparaketak egiterakoan. Sistema honek lortzen duen emaitza hobetzea izango da egiten diren esperimentuen helburua.

ontologia (*ontology*)

Mundu errealaren eskema kontzeptuala, non hitzekin izendatzen ditugun kontzeptuak modu hierarkikoan antolatuta dauden.

P@X (*precision at rank X*)

IB sistema baten eraginkortasuna neurtzeko neurri bat. Dokumentuen rankingeko X. posizioan doitasuna zenbatekoa den adierazten du.

parekatze (*matching*)

IB sistema baten prozesuetako bat, non kontsulta dokumentuen errepresentazioarekin (indizearekin) parekatu eta dokumentuen azpimultzo bat aukeratzen den.

PRF

Ikus *sasiadierazgarritasun-feedback*.

pseudo-relevance feedback

Ikus *sasiadierazgarritasun-feedback*.

QL

Ikus *kontsulta-egiantza*.

query likelihood

Ikus *kontsulta-egiantza*.

ranking-funtzio (*ranking function*)

Dokumentuen berreskurapen-prozesuan dokumentuei pisuak esleitzeko ardura duen funtzioa. IB sistema baten helburua kontsulta baten araberako dokumentuak berreskuratu eta hauek modu egokian ordenatzea da: dokumentu adierazgarriek ez-adierazgarrien aurretik egon behar dute. Hau lortzeko, sistemaren baitako ranking-funtzioak dokumentuei pisu bat esleitu behar die, eta dokumentu adierazgarriek pisu handiagoa izan beharko lukete ez-adierazgarriek baino.

sasiadierazgarritasun-feedback (*pseudo-relevance feedback*, PRF)

Kontsultaren finketa edo hedapena egiteko metodo automatiko bat. Metodo honek lehenengo berreskurapen-saiakerako lehen k dokumentuak adierazgarritzat hartuko ditu eta k dokumentu horietan oinarrituko da termino berriak lortzeko.

SemCor

WordNeteko adierekin eskuz etiketatutako ingeleseko corpus bat.

stemmer

Ikus *erro-bilatzaile*.

stopword

Dokumentuak baztertzerako garaian ekarpenik egiten ez duten eta, ondorioz, indizetik eta berreskurapen prozesutik kanpo uzten diren hitzak. Horien adibide dira, esaterako, artikuluak, preposizioak eta juntagailuak, edo bilduman oso ohikoak diren beste hainbat hitz.

synset

WordNeteko sinonimo-multzo bat (*synonim set*), kontzeptu lexikal edo adiera bati dagokiona. Hau hainbat ale lexikal edo *variant*ek osatzen dute.

testerako bilduma (*test collection*)

Datu-multzo osoaren zati bat, sistemaren ebaluazioa egiteko erabiliko dena.

***tf* (*term frequency*)**

IB arloan asko erabiltzen den neurri estatistiko bat, dokumentu batean termino batek duen agerpen kopurua adierazten duena.

tf-idf

IB arloan oso erabilia den neurketa, *tf* eta *idf* balioen biderkadura dena. Termino batentzat *tf* altua duten dokumentuak termino hori duten kontsulta batekiko adierazgarriak izango direla pentsa liteke; baina, gainera, bilduman ohikoak ez diren terminoak baztertzaileagoak dira eta informazio gehiago ematen dute asko agertzen diren terminoek

baino. Horregatik, termino bati pisua esleitzerakoan neurketa hau sarri erabili ohi da.

variant

WordNeteko *synset* bat osatzen duten ale lexikal edo hitzetako bakoitza. Hitz hauek elkarren artean sinonimoak dira. Adibidez, *car*, *auto* eta *automobile* hitzak ingeleseko WordNeteko *synset* bateko *variant*ak dira, kontzeptu bera adierazten baitute.

WordNet

Ingeleseko hitz eta adierei buruzko informazioa duen ezagutza-base lexikal bat. Izen, aditz, adjektibo eta adberbioak aurkitzen dira bertan *synset* delakoen arabera antolatuta eta hainbat erlazio semantikorekin lotuta.

A

*Stopword*en zerrenda

i	itself	has	you've
me	they	had	we've
my	them	having	they've
myself	their	do	i'd
we	theirs	does	you'd
our	themselves	did	he'd
ours	what	doing	she'd
ourselves	which	would	we'd
you	who	should	they'd
your	whom	could	'd
yours	this	ought	i'll
yourself	that	i'm	you'll
yourselves	these	'm	he'll
he	those	you're	she'll
him	am	're	we'll
his	is	he's	they'll
himself	are	's	'll
she	was	she's	isn't
her	were	it's	aren't
hers	be	we're	wasn't
herself	been	they're	weren't
it	being	i've	hasn't
its	have	've	haven't

hadn't	until	when	says
doesn't	while	where	said
don't	of	why	also
didn't	at	how	get
won't	by	all	go
wouldn't	for	any	goes
shan't	with	both	just
shouldn't	about	each	made
can't	against	few	make
cannot	between	more	put
couldn't	into	most	see
mustn't	through	other	seen
't	during	some	whether
let's	before	such	like
that's	after	no	well
who's	above	nor	back
what's	below	not	even
here's	to	only	still
there's	from	own	way
when's	up	same	take
where's	down	so	since
why's	in	than	another
how's	out	too	however
's	on	very	two
a	off	one	three
an	over	every	four
the	under	least	five
and	again	less	first
but	further	many	second
if	then	now	new
or	once	ever	old
because	here	never	high
as	there	say	long

B

Adiera-desanbiguazio bidezko hedapeneko fitxategiak

4.2 ataleko adibideei dagozkien XML fitxategiak aurki daitezke eranskin honetan. Adibideko gai eta dokumentu bakoitzarentzat HAD sistemak itzulitako XML fitxategiak eta fitxategi horien gainean egindako hedapenaren XML fitxategiak ikus daitezke. Azken fitxategi hauetan hedapeneko hitz bakoitza nondik eta nola lortu den ikusteko aukera dago.

Hona XML fitxategi horien gaineko azalpen batzuk:

- HAD sistemak itzulitako fitxategietan ("UBC sistemak desanbiguatua" eta "Lehenengo adierarekin desanbiguatua" ataletan agertzen direnak) gai edo dokumentuko hitz bakoitzarentzat <TERM> elementua dago. Elementu horretan hitzaren lemak eta kategoria gramatikalaz gain, desanbiguatu ahal izan badu, adiera bakoitzeko <SYNSET> elementua egongo da. Elementu honetan adiera horri esleitu dion pisua eta WordNeteko *synset* kodea egongo dira.
- Hedapen eta itzulpen fitxategiak aurreko fitxategiaren oso antzekoak izango dira, baina, fitxategi hauetan <SYNSET> elementuaren ondoren <EXP> elementuak (bat edo gehiago) egongo dira, baldin eta adiera hori hedatu edo itzuli bada. Hedapen edo itzulpen *osoetan* adiera guztiak egongo dira hedatuta, eta *onenetan* bakarrik TERM horrentzako SCORE handienekoa.

B.1 Gaiaren desanbiguazio, hedapen eta itzulpenak

B.1.1 Robust-WSD datu-multzoko 10.2452/064-AH gaia ingelesez (*title* eta *description* bakarrik)

Jatorrizkoa

ENtitle: Computer Mouse RSI
ENdesc: Find documents that report on computer mouse repetitive strain injuries (RSI).

UBC sistemak desanbiguatua

```
<top>
  <num>C064</num>
  <ENtitle>
    <TERM ID="C064-1" LEMA="computer" POS="NNP">
      <WF>Computer</WF>
      <SYNSET SCORE="0" CODE="07135102-n"/>
      <SYNSET SCORE="1" CODE="02481557-n"/>
    </TERM>
    <TERM ID="C064-2" LEMA="mouse" POS="NNP">
      <WF>Mouse</WF>
      <SYNSET SCORE="0" CODE="03020109-n"/>
      <SYNSET SCORE="1" CODE="01832174-n"/>
    </TERM>
    <TERM ID="C064-3" LEMA="RSI" POS="NNP">
      <WF>RSI</WF>
    </TERM>
  </ENtitle>
  <ENdesc>
    <TERM ID="C064-4" LEMA="find" POS="VBP">
      <WF>Find</WF>
      <SYNSET SCORE="0" CODE="00658116-v"/>
      <SYNSET SCORE="0" CODE="01456625-v"/>
      <SYNSET SCORE="0" CODE="00483900-v"/>
      <SYNSET SCORE="0.2212661339889 37" CODE="01538749-v"/>
      <SYNSET SCORE="0" CODE="01538351-v"/>
      <SYNSET SCORE="0" CODE="01380584-v"/>
      <SYNSET SCORE="0.19360786724032" CODE="01474694-v"/>
      <SYNSET SCORE="0" CODE="00488684-v"/>
      <SYNSET SCORE="0" CODE="01504859-v"/>
      <SYNSET SCORE="0" CODE="00364767-v"/>
      <SYNSET SCORE="0" CODE="01514923-v"/>
      <SYNSET SCORE="0.413030116779348" CODE="00622132-v"/>
      <SYNSET SCORE="0.172095881991396" CODE="01515833-v"/>
      <SYNSET SCORE="0" CODE="01124979-v"/>
      <SYNSET SCORE="0" CODE="01562037-v"/>
      <SYNSET SCORE="0" CODE="00501787-v"/>
    </TERM>
    <TERM ID="C064-5" LEMA="document" POS="NNS">
      <WF>documents</WF>
      <SYNSET SCORE="0" CODE="09653388-n"/>
      <SYNSET SCORE="0.25" CODE="02585552-n"/>
      <SYNSET SCORE="0" CODE="04886842-n"/>
      <SYNSET SCORE="0.75" CODE="04859637-n"/>
    </TERM>
    <TERM ID="C064-6" LEMA="that" POS="DT">
      <WF>that</WF>
    </TERM>
    <TERM ID="C064-7" LEMA="report" POS="NN">
      <WF>report</WF>
```

```
      <SYNSET SCORE="0" CODE="05392594-n"/>
      <SYNSET SCORE="0" CODE="04705103-n"/>
      <SYNSET SCORE="0" CODE="05009327-n"/>
      <SYNSET SCORE="0.221266133988937" CODE="05391410-n"/>
      <SYNSET SCORE="0.778733866011063" CODE="05391713-n"/>
      <SYNSET SCORE="0" CODE="04831003-n"/>
      <SYNSET SCORE="0" CODE="05502357-n"/>
    </TERM>
    <TERM ID="C064-8" LEMA="on" POS="IN">
      <WF>on</WF>
    </TERM>
    <TERM ID="C064-9" LEMA="computer" POS="NN">
      <WF>computer</WF>
      <SYNSET SCORE="0" CODE="07135102-n"/>
      <SYNSET SCORE="1" CODE="02481557-n"/>
    </TERM>
    <TERM ID="C064-10" LEMA="mouse" POS="NN">
      <WF>mouse</WF>
      <SYNSET SCORE="0" CODE="03020109-n"/>
      <SYNSET SCORE="1" CODE="01832174-n"/>
    </TERM>
    <TERM ID="C064-11" LEMA="repetitive" POS="JJ">
      <WF>repetitive</WF>
      <SYNSET SCORE="1" CODE="00560506-a"/>
      <SYNSET SCORE="0" CODE="02387175-a"/>
    </TERM>
    <TERM ID="C064-12" LEMA="strain" POS="NN">
      <WF>strain</WF>
      <SYNSET SCORE="0" CODE="00355108-n"/>
      <SYNSET SCORE="0" CODE="03852666-n"/>
      <SYNSET SCORE="0.348494161032575" CODE="07839857-n"/>
      <SYNSET SCORE="0" CODE="00411516-n"/>
      <SYNSET SCORE="0.221266133988937" CODE="10340571-n"/>
      <SYNSET SCORE="0" CODE="10314018-n"/>
      <SYNSET SCORE="0.172095881991395" CODE="05270417-n"/>
      <SYNSET SCORE="0.258143822987093" CODE="06037015-n"/>
      <SYNSET SCORE="0" CODE="04547542-n"/>
      <SYNSET SCORE="0" CODE="00505681-n"/>
      <SYNSET SCORE="0" CODE="10265132-n"/>
      <SYNSET SCORE="0" CODE="06045616-n"/>
    </TERM>
    <TERM ID="C064-13" LEMA="injury" POS="NNS">
      <WF>injuries</WF>
      <SYNSET SCORE="0.326982175783652" CODE="05451786-n"/>
      <SYNSET SCORE="0.673017824216348" CODE="10257548-n"/>
      <SYNSET SCORE="0" CODE="05469280-n"/>
      <SYNSET SCORE="0" CODE="00479770-n"/>
    </TERM>
    <TERM ID="C064-14" LEMA="(" POS="(">
      <WF>(</WF>
    </TERM>
    <TERM ID="C064-15" LEMA="RSI" POS="NNP">
      <WF>RSI</WF>
    </TERM>
    <TERM ID="C064-16" LEMA=")" POS=")">
      <WF>)</WF>
    </TERM>
    <TERM ID="C064-17" LEMA="." POS=".">
      <WF>.</WF>
    </TERM>
  </ENdesc>
  <ENnarr>
  ...
  </ENnarr>
</top>
```

Hedapen osoa

```
  <num>C064</num>
  <ENtitle>
    <TERM ID="C064-1" LEMA="computer" POS="NNP">
      <WF>Computer</WF>
```

```
      <SYNSET SCORE="0" CODE="07135102-n"><EXP>calculator</EXP><EXP>computer</EXP><EXP>
          estimator</EXP><EXP>figurer</EXP><EXP>reckoner</EXP></SYNSET>
      <SYNSET SCORE="1" CODE="02481557-n"><EXP>computer</EXP><EXP>data processor</EXP><EXP>
          electronic computer</EXP><EXP>information processing system</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-2" LEMA="mouse" POS="NNP">
      <WF>Mouse</WF>
      <SYNSET SCORE="0" CODE="03020109-n"><EXP>mouse</EXP></SYNSET>
      <SYNSET SCORE="1" CODE="01832174-n"><EXP>mouse</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-3" LEMA="RSI" POS="NNP">
      <WF>RSI</WF>
    </TERM>
  </ENtitle>
  <ENdesc>
    <TERM ID="C064-4" LEMA="find" POS="VBP">
      <WF>Find</WF>
      <SYNSET SCORE="0" CODE="00658116-v"><EXP>find</EXP><EXP>rule</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="01456625-v"><EXP>find</EXP><EXP>see</EXP><EXP>witness</EXP></
          SYNSET>
      <SYNSET SCORE="0" CODE="00483900-v"><EXP>feel</EXP><EXP>find</EXP></SYNSET>
      <SYNSET SCORE="0.221266133988937" CODE="01538749-v"><EXP>bump</EXP><EXP>chance</EXP><
          EXP>encounter</EXP><EXP>find</EXP><EXP>happen</EXP><EXP>hit</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="01538351-v"><EXP>find</EXP><EXP>recover</EXP><EXP>regain</EXP>
          <EXP>retrieve</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="01380584-v"><EXP>find</EXP></SYNSET>
      <SYNSET SCORE="0.193607867240320" CODE="01474694-v"><EXP>detect</EXP><EXP>discover</EXP
          ><EXP>find</EXP><EXP>notice</EXP><EXP>observe</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="00488684-v"><EXP>discover</EXP><EXP>find</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="01504859-v"><EXP>find</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="00364767-v"><EXP>find</EXP><EXP>get</EXP><EXP>incur</EXP><EXP>
          obtain</EXP><EXP>receive</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="01514923-v"><EXP>find</EXP></SYNSET>
      <SYNSET SCORE="0.413030116779348" CODE="00622132-v"><EXP>ascertain</EXP><EXP>determine
          </EXP><EXP>find</EXP><EXP>find out</EXP></SYNSET>
      <SYNSET SCORE="0.172095881991396" CODE="01515833-v"><EXP>come up</EXP><EXP>find</EXP><
          EXP>get hold</EXP><EXP>line up</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="01124979-v"><EXP>discover</EXP><EXP>find</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="01562037-v"><EXP>find</EXP><EXP>regain</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="00501787-v"><EXP>find</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-5" LEMA="document" POS="NNS">
      <WF>documents</WF>
      <SYNSET SCORE="0" CODE="09653388-n"><EXP>document</EXP></SYNSET>
      <SYNSET SCORE="0.25" CODE="02585552-n"><EXP>document</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="04886842-n"><EXP>document</EXP><EXP>text file</EXP></SYNSET>
      <SYNSET SCORE="0.75" CODE="04859637-n"><EXP>document</EXP><EXP>papers</EXP><EXP>
          written document</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-6" LEMA="that" POS="DT">
      <WF>that</WF>
    </TERM>
    <TERM ID="C064-7" LEMA="report" POS="NN">
      <WF>report</WF>
      <SYNSET SCORE="0" CODE="05392594-n"><EXP>report</EXP><EXP>report card</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="04705103-n"><EXP>report</EXP><EXP>reputation</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="05009327-n"><EXP>account</EXP><EXP>news report</EXP><EXP>
          report</EXP><EXP>story</EXP><EXP>write up</EXP></SYNSET>
      <SYNSET SCORE="0.221266133988937" CODE="05391410-n"><EXP>account</EXP><EXP>report</EXP
          ></SYNSET>
      <SYNSET SCORE="0.778733866011063" CODE="05391713-n"><EXP>report</EXP><EXP>study</EXP><
          /SYNSET>
      <SYNSET SCORE="0" CODE="04831003-n"><EXP>composition</EXP><EXP>paper</EXP><EXP>report<
          /EXP><EXP>theme</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="05502357-n"><EXP>report</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-8" LEMA="on" POS="IN">
      <WF>on</WF>
    </TERM>
    <TERM ID="C064-9" LEMA="computer" POS="NN">
      <WF>computer</WF>
      <SYNSET SCORE="0" CODE="07135102-n"><EXP>calculator</EXP><EXP>computer</EXP><EXP>
          estimator</EXP><EXP>figurer</EXP><EXP>reckoner</EXP></SYNSET>
      <SYNSET SCORE="1" CODE="02481557-n"><EXP>computer</EXP><EXP>data processor</EXP><EXP>
          electronic computer</EXP><EXP>information processing system</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-10" LEMA="mouse" POS="NN">
```

```
      <WF>mouse</WF>
      <SYNSET SCORE="0" CODE="03020109-n"><EXP>mouse</EXP></SYNSET>
      <SYNSET SCORE="1" CODE="01832174-n"><EXP>mouse</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-11" LEMA="repetitive" POS="JJ">
      <WF>repetitive</WF>
      <SYNSET SCORE="1" CODE="00560506-a"><EXP>insistent</EXP><EXP>repetitive</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="02387175-a"><EXP>iterative</EXP><EXP>reiterative</EXP><EXP>
          repetitious</EXP><EXP>repetitive</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-12" LEMA="strain" POS="NN">
      <WF>strain</WF>
      <SYNSET SCORE="0" CODE="00355108-n"><EXP>song</EXP><EXP>strain</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="03852666-n"><EXP>breed</EXP><EXP>strain</EXP></SYNSET>
      <SYNSET SCORE="0.348494161032575" CODE="07839857-n"><EXP>strain</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="00411516-n"><EXP>strain</EXP><EXP>straining</EXP><EXP>stress</
          EXP></SYNSET>
      <SYNSET SCORE="0.221266133988937" CODE="10340571-n"><EXP>strain</EXP><EXP>stress</EXP>
          </SYNSET>
      <SYNSET SCORE="0" CODE="10314018-n"><EXP>mental strain</EXP><EXP>nervous strain</EXP><
          EXP>strain</EXP></SYNSET>
      <SYNSET SCORE="0.172095881991395" CODE="05270417-n"><EXP>air</EXP><EXP>line</EXP><EXP>
          melodic line</EXP><EXP>melodic phrase</EXP><EXP>melody</EXP><EXP>strain</EXP><EXP
          >tune</EXP></SYNSET>
      <SYNSET SCORE="0.258143822987093" CODE="06037015-n"><EXP>breed</EXP><EXP>stock</EXP><
          EXP>strain</EXP><EXP>variety</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="04547542-n"><EXP>strain</EXP><EXP>tenor</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="00505681-n"><EXP>nisus</EXP><EXP>pains</EXP><EXP>strain</EXP><
          EXP>striving</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="10265132-n"><EXP>strain</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="06045616-n"><EXP>form</EXP><EXP>strain</EXP><EXP>var.</EXP><
          EXP>variant</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-13" LEMA="injury" POS="NNS">
      <WF>injuries</WF>
      <SYNSET SCORE="0.326982175783652" CODE="05451786-n"><EXP>accidental injury</EXP><EXP>
          injury</EXP></SYNSET>
      <SYNSET SCORE="0.673017824216348" CODE="10257548-n"><EXP>harm</EXP><EXP>hurt</EXP><EXP
          >injury</EXP><EXP>trauma</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="05469280-n"><EXP>combat injury</EXP><EXP>injury</EXP><EXP>
          wound</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="00479770-n"><EXP>injury</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-14" LEMA="(" POS="(">
      <WF>(</WF>
    </TERM>
    <TERM ID="C064-15" LEMA="RSI" POS="NNP">
      <WF>RSI</WF>
    </TERM>
    <TERM ID="C064-16" LEMA=")" POS=")">
      <WF>)</WF>
    </TERM>
    <TERM ID="C064-17" LEMA="." POS=".">
      <WF>.</WF>
    </TERM>
  </ENdesc>
  <ENnarr>
  ...
  </ENnarr>
</top>
```

Hedapen onena

```
<top>
  <num>C064</num>
  <ENtitle>
    <TERM ID="C064-1" LEMA="computer" POS="NNP">
      <WF>Computer</WF>
      <SYNSET SCORE="0" CODE="07135102-n"/>
      <SYNSET SCORE="1" CODE="02481557-n"><EXP>computer</EXP><EXP>data processor</EXP><EXP>
          electronic computer</EXP><EXP>information processing system</EXP></SYNSET>
    </TERM>
```

```
    <TERM ID="C064-2" LEMA="mouse" POS="NNP">
      <WF>Mouse</WF>
      <SYNSET SCORE="0" CODE="03020109-n"/>
      <SYNSET SCORE="1" CODE="01832174-n"><EXP>mouse</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-3" LEMA="RSI" POS="NNP">
      <WF>RSI</WF>
    </TERM>
  </ENtitle>
  <ENdesc>
    <TERM ID="C064-4" LEMA="find" POS="VBP">
      <WF>Find</WF>
      <SYNSET SCORE="0" CODE="00658116-v"/>
      <SYNSET SCORE="0" CODE="01456625-v"/>
      <SYNSET SCORE="0" CODE="00483900-v"/>
      <SYNSET SCORE="0.221266133988937" CODE="01538749-v"/>
      <SYNSET SCORE="0" CODE="01538351-v"/>
      <SYNSET SCORE="0" CODE="01380584-v"/>
      <SYNSET SCORE="0.19360786724032" CODE="01474694-v"/>
      <SYNSET SCORE="0" CODE="00488684-v"/>
      <SYNSET SCORE="0" CODE="01504859-v"/>
      <SYNSET SCORE="0" CODE="00364767-v"/>
      <SYNSET SCORE="0" CODE="01514923-v"/>
      <SYNSET SCORE="0.413030116779348" CODE="00622132-v"><EXP>ascertain</EXP><EXP>determine
          </EXP><EXP>find</EXP><EXP>find out</EXP></SYNSET>
      <SYNSET SCORE="0.172095881991396" CODE="01515833-v"/>
      <SYNSET SCORE="0" CODE="01124979-v"/>
      <SYNSET SCORE="0" CODE="01562037-v"/>
      <SYNSET SCORE="0" CODE="00501787-v"/>
    </TERM>
    <TERM ID="C064-5" LEMA="document" POS="NNS">
      <WF>documents</WF>
      <SYNSET SCORE="0" CODE="09653388-n"/>
      <SYNSET SCORE="0.25" CODE="02585552-n"/>
      <SYNSET SCORE="0" CODE="04886842-n"/>
      <SYNSET SCORE="0.75" CODE="04859637-n"><EXP>document</EXP><EXP>papers</EXP><EXP>
          written document</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-6" LEMA="that" POS="DT">
      <WF>that</WF>
    </TERM>
    <TERM ID="C064-7" LEMA="report" POS="NN">
      <WF>report</WF>
      <SYNSET SCORE="0" CODE="05392594-n"/>
      <SYNSET SCORE="0" CODE="04705103-n"/>
      <SYNSET SCORE="0" CODE="05009327-n"/>
      <SYNSET SCORE="0.221266133988937" CODE="05391410-n"/>
      <SYNSET SCORE="0.778733866011063" CODE="05391713-n"><EXP>report</EXP><EXP>study</EXP><
          /SYNSET>
      <SYNSET SCORE="0" CODE="04831003-n"/>
      <SYNSET SCORE="0" CODE="05502357-n"/>
    </TERM>
    <TERM ID="C064-8" LEMA="on" POS="IN">
      <WF>on</WF>
    </TERM>
    <TERM ID="C064-9" LEMA="computer" POS="NN">
      <WF>computer</WF>
      <SYNSET SCORE="0" CODE="07135102-n"/>
      <SYNSET SCORE="1" CODE="02481557-n"><EXP>computer</EXP><EXP>data processor</EXP><EXP>
          electronic computer</EXP><EXP>information processing system</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-10" LEMA="mouse" POS="NN">
      <WF>mouse</WF>
      <SYNSET SCORE="0" CODE="03020109-n"/>
      <SYNSET SCORE="1" CODE="01832174-n"><EXP>mouse</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-11" LEMA="repetitive" POS="JJ">
      <WF>repetitive</WF>
      <SYNSET SCORE="1" CODE="00560506-a"><EXP>insistent</EXP><EXP>repetitive</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="02387175-a"/>
    </TERM>
    <TERM ID="C064-12" LEMA="strain" POS="NN">
      <WF>strain</WF>
      <SYNSET SCORE="0" CODE="00355108-n"/>
      <SYNSET SCORE="0" CODE="03852666-n"/>
      <SYNSET SCORE="0.348494161032575" CODE="07839857-n"><EXP>strain</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="00411516-n"/>
```

```
      <SYNSET SCORE="0.221266133988937" CODE="10340571-n"/>
      <SYNSET SCORE="0" CODE="10314018-n"/>
      <SYNSET SCORE="0.172095881991395" CODE="05270417-n"/>
      <SYNSET SCORE="0.258143822987093" CODE="06037015-n"/>
      <SYNSET SCORE="0" CODE="04547542-n"/>
      <SYNSET SCORE="0" CODE="00505681-n"/>
      <SYNSET SCORE="0" CODE="10265132-n"/>
      <SYNSET SCORE="0" CODE="06045616-n"/>
    </TERM>
    <TERM ID="C064-13" LEMA="injury" POS="NNS">
      <WF>injuries</WF>
      <SYNSET SCORE="0.326982175783652" CODE="05451786-n"/>
      <SYNSET SCORE="0.673017824216348" CODE="10257548-n"><EXP>harm</EXP><EXP>hurt</EXP><EXP
          >injury</EXP><EXP>trauma</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="05469280-n"/>
      <SYNSET SCORE="0" CODE="00479770-n"/>
    </TERM>
    <TERM ID="C064-14" LEMA="(" POS="(">
      <WF>(</WF>
    </TERM>
    <TERM ID="C064-15" LEMA="RSI" POS="NNP">
      <WF>RSI</WF>
    </TERM>
    <TERM ID="C064-16" LEMA=")" POS=")">
      <WF>)</WF>
    </TERM>
    <TERM ID="C064-17" LEMA="." POS=".">
      <WF>.</WF>
    </TERM>
  </ENdesc>
  <ENnarr>
  ...
  </ENnarr>
</top>
```

B.1.2 Robust-WSD datu-multzoko 10.2452/064-AH gaia gaztelaniaz (*title* eta *description* bakarrik)

Jatorrizkoa

EStitle: Síndrome RSI y ratones de ordenador
ESdesc: Encontrar documentos que informen sobre RSI ("repetitive strain injuries" o "enfermedad del periodista") producidas por el uso del ratón del ordenador.

Lehenengo adierarekin desanbiguatua

```
<top>
  <num>C064</num>
  <EStitle>
    <TERM ID="C064-1" LEMA="sindrome_rsi" POS="NP00000">
      <WF>Sindrome_RSI</WF>
    </TERM>
    <TERM ID="C064-2" LEMA="y" POS="CC">
      <WF>y</WF>
    </TERM>
    <TERM ID="C064-3" LEMA="raton" POS="NCMP000">
      <WF>ratones</WF>
      <SYNSET SCORE="1" CODE="01832174-n"/>
    </TERM>
    <TERM ID="C064-4" LEMA="de" POS="SPS00">
      <WF>de</WF>
    </TERM>
    <TERM ID="C064-5" LEMA="ordenador" POS="NCMS000">
      <WF>ordenador</WF>
```

```
      <SYNSET SCORE="1" CODE="02481557-n"/>
    </TERM>
  </ESstitle>
  <ESdesc>
    <TERM ID="C064-6" LEMA="encontrar" POS="VMN0000">
      <WF>Encontrar</WF>
      <SYNSET SCORE="1" CODE="01474694-v"/>
    </TERM>
    <TERM ID="C064-7" LEMA="documento" POS="NCMP000">
      <WF>documentos</WF>
      <SYNSET SCORE="1" CODE="04859637-n"/>
    </TERM>
    <TERM ID="C064-8" LEMA="que" POS="PR0CN000">
      <WF>que</WF>
    </TERM>
    <TERM ID="C064-9" LEMA="informar" POS="VMSP3P0">
      <WF>informen</WF>
      <SYNSET SCORE="1" CODE="00655029-v"/>
    </TERM>
    <TERM ID="C064-10" LEMA="sobre" POS="SPS00">
      <WF>sobre</WF>
    </TERM>
    <TERM ID="C064-11" LEMA="rsi" POS="NP00000">
      <WF>RSI</WF>
      </TERM>
    <TERM ID="C064-12" LEMA="(" POS="Fpa">
      <WF>(</WF>
    </TERM>
    <TERM ID="C064-13" LEMA="&amp;quot;" POS="Fe">
      <WF>&amp;quot;</WF>
    </TERM>
    <TERM ID="C064-14" LEMA="repetitive" POS="VMIP3S0">
      <WF>repetitive</WF>
    </TERM>
    <TERM ID="C064-15" LEMA="strain" POS="NP00000">
      <WF>strain</WF>
    </TERM>
    <TERM ID="C064-16" LEMA="injuriar" POS="VMSP2S0">
      <WF>injuries</WF>
      <SYNSET SCORE="1" CODE="00572942-v"/>
    </TERM>
    <TERM ID="C064-17" LEMA="&amp;quot;" POS="Fe">
      <WF>&amp;quot;</WF>
    </TERM>
    <TERM ID="C064-18" LEMA="o" POS="CC">
      <WF>o</WF>
    </TERM>
    <TERM ID="C064-19" LEMA="&amp;quot;" POS="Fe">
      <WF>&amp;quot;</WF>
    </TERM>
    <TERM ID="C064-20" LEMA="enfermedad" POS="NCFS000">
      <WF>enfermedad</WF>
      <SYNSET SCORE="1" CODE="10129713-n"/>
    </TERM>
    <TERM ID="C064-21" LEMA="de" POS="SPS00">
      <WF>de</WF>
    </TERM>
    <TERM ID="C064-22" LEMA="el" POS="DA0MS0">
      <WF>el</WF>
    </TERM>
    <TERM ID="C064-23" LEMA="periodista" POS="NCCS000">
      <WF>periodista</WF>
      <SYNSET SCORE="1" CODE="07528776-n"/>
    </TERM>
    <TERM ID="C064-24" LEMA="&amp;quot;" POS="Fe">
      <WF>&amp;quot;</WF>
    </TERM>
    <TERM ID="C064-25" LEMA=")" POS="Fpt">
      <WF>)</WF>
    </TERM>
    <TERM ID="C064-26" LEMA="producir" POS="VMP00PF">
      <WF>producidas</WF>
      <SYNSET SCORE="1" CODE="01114991-v"/>
    </TERM>
    <TERM ID="C064-27" LEMA="por" POS="SPS00">
      <WF>por</WF>
    </TERM>
```

```
    <TERM ID="C064-28" LEMA="el" POS="DA0MS0">
      <WF>el</WF>
    </TERM>
    <TERM ID="C064-29" LEMA="uso" POS="NCMS000">
      <WF>uso</WF>
      <SYNSET SCORE="1" CODE="00605730-n"/>
    </TERM>
    <TERM ID="C064-30" LEMA="de" POS="SPS00">
      <WF>de</WF>
    </TERM>
    <TERM ID="C064-31" LEMA="el" POS="DA0MS0">
      <WF>el</WF>
    </TERM>
    <TERM ID="C064-32" LEMA="raton" POS="NCMS000">
      <WF>raton</WF>
      <SYNSET SCORE="1" CODE="01832174-n"/>
    </TERM>
    <TERM ID="C064-33" LEMA="de" POS="SPS00">
      <WF>de</WF>
    </TERM>
    <TERM ID="C064-34" LEMA="el" POS="DA0MS0">
      <WF>el</WF>
    </TERM>
    <TERM ID="C064-35" LEMA="ordenador" POS="NCMS000">
      <WF>ordenador</WF>
      <SYNSET SCORE="1" CODE="02481557-n"/>
    </TERM>
    <TERM ID="C064-36" LEMA="." POS="Fp">
      <WF>.</WF>
    </TERM>
  </ESdesc>
  <ESnarr>
  ...
  </ESnarr>
</top>
```

Itzulpena ingelesera

```
<top>
  <num> C064 </num>
  <ESTitle>
    <TERM ID="C064-1" LEMA="sindrome_rsi" POS="NP00000">
      <WF>Sindrome_RSI</WF>
    </TERM>
    <TERM ID="C064-2" LEMA="y" POS="CC">
      <WF>y</WF>
    </TERM>
    <TERM ID="C064-3" LEMA="raton" POS="NCMP000">
      <WF>ratones</WF>
      <SYNSET SCORE="1" CODE="01832174-n"><EXP>mouse</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-4" LEMA="de" POS="SPS00">
      <WF>de</WF>
    </TERM>
    <TERM ID="C064-5" LEMA="ordenador" POS="NCMS000">
      <WF>ordenador</WF>
      <SYNSET SCORE="1" CODE="02481557-n"><EXP>computer</EXP><EXP>data processor</EXP><EXP>
          electronic computer</EXP><EXP>information processing system</EXP></SYNSET>
    </TERM>
  </ESTitle>
  <ESdesc>
    <TERM ID="C064-6" LEMA="encontrar" POS="VMN0000">
      <WF>Encontrar</WF>
      <SYNSET SCORE="1" CODE="01474694-v"><EXP>detect</EXP><EXP>discover</EXP><EXP>find</EXP
          ><EXP>notice</EXP><EXP>observe</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-7" LEMA="documento" POS="NCMP000">
      <WF>documentos</WF>
      <SYNSET SCORE="1" CODE="04859637-n"><EXP>document</EXP><EXP>papers</EXP><EXP>written
          document</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-8" LEMA="que" POS="PR0CN000">
```

```
    <WF>que</WF>
  </TERM>
  <TERM ID="C064-9" LEMA="informar" POS="VMSP3P0">
    <WF>informen</WF>
    <SYNSET SCORE="1" CODE="00655029-v"><EXP>report</EXP></SYNSET>
  </TERM>
  <TERM ID="C064-10" LEMA="sobre" POS="SPS00">
    <WF>sobre</WF>
  </TERM>
  <TERM ID="C064-11" LEMA="rsi" POS="NP00000">
    <WF>RSI</WF>
  </TERM>
  <TERM ID="C064-12" LEMA="(" POS="Fpa">
    <WF>(</WF>
  </TERM>
  <TERM ID="C064-13" LEMA="&amp;quot;" POS="Fe">
    <WF>''</WF>
  </TERM>
  <TERM ID="C064-14" LEMA="repetitive" POS="VMIP3S0">
    <WF>repetitive</WF>
  </TERM>
  <TERM ID="C064-15" LEMA="strain" POS="NP00000">
    <WF>strain</WF>
  </TERM>
  <TERM ID="C064-16" LEMA="injuriar" POS="VMSP2S0">
    <WF>injuries</WF>
    <SYNSET SCORE="1" CODE="00572942-v"><EXP>abuse</EXP><EXP>blackguard</EXP><EXP>
        clapperclaw</EXP><EXP>shout</EXP></SYNSET>
  </TERM>
  <TERM ID="C064-17" LEMA="&amp;quot;" POS="Fe">
    <WF>''</WF>
  </TERM>
  <TERM ID="C064-18" LEMA="o" POS="CC">
    <WF>o</WF>
  </TERM>
  <TERM ID="C064-19" LEMA="&amp;quot;" POS="Fe">
    <WF>''</WF>
  </TERM>
  <TERM ID="C064-20" LEMA="enfermedad" POS="NCFS000">
    <WF>enfermedad</WF>
    <SYNSET SCORE="1" CODE="10129713-n"><EXP>disease</EXP></SYNSET>
  </TERM>
  <TERM ID="C064-21" LEMA="de" POS="SPS00">
    <WF>de</WF>
  </TERM>
  <TERM ID="C064-22" LEMA="el" POS="DA0MS0">
    <WF>el</WF>
  </TERM>
  <TERM ID="C064-23" LEMA="periodista" POS="NCCS000">
    <WF>periodista</WF>
    <SYNSET SCORE="1" CODE="07528776-n"><EXP>newsman</EXP><EXP>newsperson</EXP><EXP>
        reporter</EXP></SYNSET>
  </TERM>
  <TERM ID="C064-24" LEMA="&amp;quot;" POS="Fe">
    <WF>''</WF>
  </TERM>
  <TERM ID="C064-25" LEMA=")" POS="Fpt">
    <WF>)</WF>
  </TERM>
  <TERM ID="C064-26" LEMA="producir" POS="VMP00PF">
    <WF>producidas</WF>
    <SYNSET SCORE="1" CODE="01114991-v"><EXP>create</EXP><EXP>make</EXP><EXP>produce</EXP>
      </SYNSET>
  </TERM>
  <TERM ID="C064-27" LEMA="por" POS="SPS00">
    <WF>por</WF>
  </TERM>
  <TERM ID="C064-28" LEMA="el" POS="DA0MS0">
    <WF>el</WF>
  </TERM>
  <TERM ID="C064-29" LEMA="uso" POS="NCMS000">
    <WF>uso</WF>
    <SYNSET SCORE="1" CODE="00605730-n"><EXP>employment</EXP><EXP>exercise</EXP><EXP>usage
      </EXP><EXP>use</EXP><EXP>utilisation</EXP><EXP>utilization</EXP></SYNSET>
  </TERM>
  <TERM ID="C064-30" LEMA="de" POS="SPS00">
    <WF>de</WF>
```

```
    </TERM>
    <TERM ID="C064-31" LEMA="el" POS="DA0MS0">
      <WF>el</WF>
    </TERM>
    <TERM ID="C064-32" LEMA="raton" POS="NCMS000">
      <WF>raton</WF>
      <SYNSET SCORE="1" CODE="01832174-n"><EXP>mouse</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-33" LEMA="de" POS="SPS00">
      <WF>de</WF>
    </TERM>
    <TERM ID="C064-34" LEMA="el" POS="DA0MS0">
      <WF>el</WF>
    </TERM>
    <TERM ID="C064-35" LEMA="ordenador" POS="NCMS000">
      <WF>ordenador</WF>
      <SYNSET SCORE="1" CODE="02481557-n"><EXP>computer</EXP><EXP>data processor</EXP><EXP>
          electronic computer</EXP><EXP>information processing system</EXP></SYNSET>
    </TERM>
    <TERM ID="C064-36" LEMA="." POS="Fp">
      <WF>.</WF>
    </TERM>
  </ESdesc>
  <ESnarr>
  ...
  </ESnarr>
</top>
```

B.2 Dokumentuaren desanbiguazio eta itzulpenak

B.2.1 Robust-WSD datu-multzoko LA081794-0225 dokumentua ingelesez (titularra bakarrik)

Jatorrizkoa

2 FIRMS ADOPT LABELS WARNING COMPUTER USERS ABOUT DANGER OF INJURY; TECHNOLOGY: COMPAQ, MICROSOFT ARE THE FIRST TO STATE THAT HARM COULD COME FROM KEYBOARD MISUSE OR TOO MUCH TYPING.

UBC sistemak desanbiguatua

```
<DOC>
  <DOCNO>LA081794-0225</DOCNO>
  <DOCID>LA081794-0225</DOCID>
  <HEADLINE>
    <TERM ID="LA081794-0225-1" LEMA="2" POS="CD">
      <WF>2</WF>
    </TERM>
    <TERM ID="LA081794-0225-2" LEMA="firm" POS="NNPS">
      <WF>FIRMS</WF>
      <SYNSET SCORE="1" CODE="06007316-n"/>
    </TERM>
    <TERM ID="LA081794-0225-3" LEMA="ADOPT" POS="NNP">
      <WF>ADOPT</WF>
    </TERM>
    <TERM ID="LA081794-0225-4" LEMA="label" POS="NNP">
      <WF>LABELS</WF>
      <SYNSET SCORE="0.428571428571429" CODE="05381676-n"/>
      <SYNSET SCORE="0.142857142857143" CODE="05427590-n"/>
```

```
    <SYNSET SCORE="0.285714285714286" CODE="05128468-n"/>
    <SYNSET SCORE="0.142857142857143" CODE="10472538-n"/>
  </TERM>
  <TERM ID="LA081794-0225-5" LEMA="warn" POS="VBG">
    <WF>WARNING</WF>
    <SYNSET SCORE="1" CODE="00589833-v"/>
    <SYNSET SCORE="0" CODE="00590070-v"/>
  </TERM>
  <TERM ID="LA081794-0225-6" LEMA="computer" POS="NNP">
    <WF>COMPUTER</WF>
    <SYNSET SCORE="0" CODE="07135102-n"/>
    <SYNSET SCORE="1" CODE="02481557-n"/>
  </TERM>
  <TERM ID="LA081794-0225-7" LEMA="user" POS="NNPS">
    <WF>USERS</WF>
    <SYNSET SCORE="0.25" CODE="07252689-n"/>
    <SYNSET SCORE="0" CODE="07229219-n"/>
    <SYNSET SCORE="0.75" CODE="07658488-n"/>
  </TERM>
  <TERM ID="LA081794-0225-8" LEMA="ABOUT" POS="NNP">
    <WF>ABOUT</WF>
  </TERM>
  <TERM ID="LA081794-0225-9" LEMA="danger" POS="NNP">
    <WF>DANGER</WF>
    <SYNSET SCORE="0.585125998770744" CODE="00512734-n"/>
    <SYNSET SCORE="0" CODE="06365821-n"/>
    <SYNSET SCORE="0.221266133988937" CODE="10427605-n"/>
    <SYNSET SCORE="0.19360786724032" CODE="10427326-n"/>
  </TERM>
  <TERM ID="LA081794-0225-10" LEMA="OF" POS="IN">
    <WF>OF</WF>
  </TERM>
  <TERM ID="LA081794-0225-11" LEMA="injury" POS="NN">
    <WF>INJURY</WF>
    <SYNSET SCORE="0.569760295021511" CODE="10257548-n"/>
    <SYNSET SCORE="0.430239704978489" CODE="05451786-n"/>
    <SYNSET SCORE="0" CODE="00479770-n"/>
    <SYNSET SCORE="0" CODE="05469280-n"/>
  </TERM>
  <TERM ID="LA081794-0225-12" LEMA=";" POS=":">
    <WF>;</WF>
  </TERM>
  <TERM ID="LA081794-0225-13" LEMA="technology" POS="NNP">
    <WF>TECHNOLOGY</WF>
    <SYNSET SCORE="0.827904118008605" CODE="00607693-n"/>
    <SYNSET SCORE="0.172095881991395" CODE="04660658-n"/>
  </TERM>
  <TERM ID="LA081794-0225-14" LEMA=":" POS=":">
    <WF>:</WF>
  </TERM>
  <TERM ID="LA081794-0225-15" LEMA="COMPAQ" POS="NNP">
    <WF>COMPAQ</WF>
  </TERM>
  <TERM ID="LA081794-0225-16" LEMA="," POS=",">
    <WF>,</WF>
  </TERM>
  <TERM ID="LA081794-0225-17" LEMA="MICROSOFT" POS="NNP">
    <WF>MICROSOFT</WF>
  </TERM>
  <TERM ID="LA081794-0225-18" LEMA="be" POS="VBP">
    <WF>ARE</WF>
    <SYNSET SCORE="0.0113384126222329" CODE="01787769-v"/>
    <SYNSET SCORE="0.181174635551023" CODE="01784339-v"/>
    <SYNSET SCORE="0.644489771431999" CODE="01775973-v"/>
    <SYNSET SCORE="0.00515927770112184" CODE="01666138-v"/>
    <SYNSET SCORE="0.0420541124242606" CODE="01775163-v"/>
    <SYNSET SCORE="0.00347951286819845" CODE="01840295-v"/>
    <SYNSET SCORE="0.0540524326594277" CODE="01811792-v"/>
    <SYNSET SCORE="0" CODE="01843641-v"/>
    <SYNSET SCORE="0.000119983202351671" CODE="01552250-v"/>
    <SYNSET SCORE="0.0418741376207331" CODE="01781222-v"/>
    <SYNSET SCORE="5.99916011758354e-05" CODE="01782836-v"/>
    <SYNSET SCORE="0.0161977323174756" CODE="01817610-v"/>
  </TERM>
  <TERM ID="LA081794-0225-19" LEMA="THE" POS="DT">
    <WF>THE</WF>
  </TERM>
```

```
    <TERM ID="LA081794-0225-20" LEMA="first" POS="NNP">
      <WF>FIRST</WF>
      <SYNSET SCORE="1" CODE="09974999-n"/>
      <SYNSET SCORE="0" CODE="09770029-n"/>
      <SYNSET SCORE="0" CODE="00464759-n"/>
      <SYNSET SCORE="0" CODE="10965545-n"/>
      <SYNSET SCORE="0" CODE="02688619-n"/>
      <SYNSET SCORE="0" CODE="05023676-n"/>
    </TERM>
    <TERM ID="LA081794-0225-21" LEMA="TO" POS="TO">
      <WF>TO</WF>
    </TERM>
    <TERM ID="LA081794-0225-22" LEMA="state" POS="NNP">
      <WF>STATE</WF>
      <SYNSET SCORE="0" CODE="06060831-n"/>
      <SYNSET SCORE="0" CODE="06374245-n"/>
      <SYNSET SCORE="0.376152427781192" CODE="06074189-n"/>
      <SYNSET SCORE="0.451751690227412" CODE="06079469-n"/>
      <SYNSET SCORE="0.172095881991396" CODE="00016185-n"/>
      <SYNSET SCORE="0" CODE="10077290-n"/>
      <SYNSET SCORE="0" CODE="06299747-n"/>
      <SYNSET SCORE="0" CODE="10386919-n"/>
    </TERM>
    <TERM ID="LA081794-0225-23" LEMA="THAT" POS="WDT">
      <WF>THAT</WF>
    </TERM>
    <TERM ID="LA081794-0225-24" LEMA="harm" POS="NNP">
      <WF>HARM</WF>
      <SYNSET SCORE="0.25" CODE="05522654-n"/>
      <SYNSET SCORE="0.625" CODE="10257548-n"/>
      <SYNSET SCORE="0.125" CODE="00258668-n"/>
    </TERM>
    <TERM ID="LA081794-0225-25" LEMA="COULD" POS="NNP">
      <WF>COULD</WF>
    </TERM>
    <TERM ID="LA081794-0225-26" LEMA="COME" POS="NNP">
      <WF>COME</WF>
    </TERM>
    <TERM ID="LA081794-0225-27" LEMA="FROM" POS="NNP">
      <WF>FROM</WF>
    </TERM>
    <TERM ID="LA081794-0225-28" LEMA="keyboard" POS="NNP">
      <WF>KEYBOARD</WF>
      <SYNSET SCORE="1" CODE="02887166-n"/>
      <SYNSET SCORE="0" CODE="02887056-n"/>
    </TERM>
    <TERM ID="LA081794-0225-29" LEMA="misuse" POS="NNP">
      <WF>MISUSE</WF>
      <SYNSET SCORE="1" CODE="00606110-n"/>
    </TERM>
    <TERM ID="LA081794-0225-30" LEMA="OR" POS="NNP">
      <WF>OR</WF>
    </TERM>
    <TERM ID="LA081794-0225-31" LEMA="TOO" POS="NNP">
      <WF>TOO</WF>
    </TERM>
    <TERM ID="LA081794-0225-32" LEMA="much" POS="NNP">
      <WF>MUCH</WF>
      <SYNSET SCORE="1" CODE="09921720-n"/>
    </TERM>
    <TERM ID="LA081794-0225-33" LEMA="type" POS="VBG">
      <WF>TYPING</WF>
      <SYNSET SCORE="0.333333333333333" CODE="00418418-v"/>
      <SYNSET SCORE="0.666666666666667" CODE="00679313-v"/>
    </TERM>
    <TERM ID="LA081794-0225-34" LEMA="." POS=".">
      <WF>.</WF>
    </TERM>
  </HEADLINE>
  <TEXT>
  ...
  </TEXT>
</DOC>
```

Itzulpen osoa gaztelaniara

```
<DOC>
  <DOCNO>LA081794-0225</DOCNO>
  <DOCID>LA081794-0225</DOCID>
  <HEADLINE>
    <TERM ID="LA081794-0225-1" LEMA="2" POS="CD">
      <WF>2</WF>
    </TERM>
    <TERM ID="LA081794-0225-2" LEMA="firm" POS="NNPS">
      <WF>FIRMS</WF>
      <SYNSET SCORE="1" CODE="06007316-n"/>
    </TERM>
    <TERM ID="LA081794-0225-3" LEMA="ADOPT" POS="NNP">
      <WF>ADOPT</WF>
    </TERM>
    <TERM ID="LA081794-0225-4" LEMA="label" POS="NNP">
      <WF>LABELS</WF>
      <SYNSET SCORE="0.428571428571429" CODE="05381676-n"><EXP>etiqueta</EXP></SYNSET>
      <SYNSET SCORE="0.142857142857143" CODE="05427590-n"><EXP>etiqueta</EXP><EXP>pegatina</
          EXP><EXP>pegata</EXP><EXP>etiqueta adhesiva</EXP></SYNSET>
      <SYNSET SCORE="0.285714285714286" CODE="05128468-n"><EXP>marca</EXP></SYNSET>
      <SYNSET SCORE="0.142857142857143" CODE="10472538-n"/>
    </TERM>
    <TERM ID="LA081794-0225-5" LEMA="warn" POS="VBG">
      <WF>WARNING</WF>
      <SYNSET SCORE="1" CODE="00589833-v"><EXP>avisar</EXP><EXP>notificar</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="00590070-v"><EXP>amonestar</EXP><EXP>prevenir</EXP><EXP>avisar
          </EXP></SYNSET>
    </TERM>
    <TERM ID="LA081794-0225-6" LEMA="computer" POS="NNP">
      <WF>COMPUTER</WF>
      <SYNSET SCORE="0" CODE="07135102-n"/>
      <SYNSET SCORE="1" CODE="02481557-n"><EXP>ordenador</EXP><EXP>procesador</EXP></SYNSET>
    </TERM>
    <TERM ID="LA081794-0225-7" LEMA="user" POS="NNPS">
      <WF>USERS</WF>
      <SYNSET SCORE="0.25" CODE="07252689-n"><EXP>explotador</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="07229219-n"><EXP>consumidor de drogas</EXP></SYNSET>
      <SYNSET SCORE="0.75" CODE="07658488-n"><EXP>usuario</EXP></SYNSET>
    </TERM>
    <TERM ID="LA081794-0225-8" LEMA="ABOUT" POS="NNP">
      <WF>ABOUT</WF>
    </TERM>
    <TERM ID="LA081794-0225-9" LEMA="danger" POS="NNP">
      <WF>DANGER</WF>
      <SYNSET SCORE="0.585125998770744" CODE="00512734-n"><EXP>peligro</EXP><EXP>riesgo</EXP
          ></SYNSET>
      <SYNSET SCORE="0" CODE="06365821-n"/>
      <SYNSET SCORE="0.221266133988937" CODE="10427605-n"><EXP>peligro</EXP></SYNSET>
      <SYNSET SCORE="0.19360786724032" CODE="10427326-n"><EXP>peligro</EXP></SYNSET>
    </TERM>
    <TERM ID="LA081794-0225-10" LEMA="OF" POS="IN">
      <WF>OF</WF>
    </TERM>
    <TERM ID="LA081794-0225-11" LEMA="injury" POS="NN">
      <WF>INJURY</WF>
       <SYNSET SCORE="0.569760295021511" CODE="10257548-n"><EXP>trauma</EXP><EXP>dano</EXP><
           EXP>contusion</EXP><EXP>herida</EXP><EXP>lesion</EXP><EXP>traumatismo</EXP></
           SYNSET>
      <SYNSET SCORE="0.430239704978489" CODE="05451786-n"><EXP>herida accidental</EXP></
          SYNSET>
      <SYNSET SCORE="0" CODE="00479770-n"><EXP>agravio</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="05469280-n"><EXP>herida</EXP><EXP>herida de guerra</EXP></
          SYNSET>
    </TERM>
    <TERM ID="LA081794-0225-12" LEMA=";" POS=":">
      <WF>;</WF>
    </TERM>
    <TERM ID="LA081794-0225-13" LEMA="technology" POS="NNP">
      <WF>TECHNOLOGY</WF>
      <SYNSET SCORE="0.827904118008605" CODE="00607693-n"><EXP>ingenieria</EXP><EXP>
          tecnologia</EXP></SYNSET>
```

```
  <SYNSET SCORE="0.172095881991395" CODE="04660658-n"><EXP>ciencia aplicada</EXP><EXP>
      tecnologia</EXP><EXP>ingenieria</EXP></SYNSET>
</TERM>
<TERM ID="LA081794-0225-14" LEMA=":" POS=":">
  <WF>:</WF>
</TERM>
<TERM ID="LA081794-0225-15" LEMA="COMPAQ" POS="NNP">
  <WF>COMPAQ</WF>
</TERM>
<TERM ID="LA081794-0225-16" LEMA="," POS=",">
  <WF>,</WF>
</TERM>
<TERM ID="LA081794-0225-17" LEMA="MICROSOFT" POS="NNP">
  <WF>MICROSOFT</WF>
</TERM>
<TERM ID="LA081794-0225-18" LEMA="be" POS="VBP">
  <WF>ARE</WF>
  <SYNSET SCORE="0.0113384126222329" CODE="01787769-v"><EXP>representar</EXP><EXP>ser</
      EXP></SYNSET>
  <SYNSET SCORE="0.181174635551023" CODE="01784339-v"/>
  <SYNSET SCORE="0.644489771431999" CODE="01775973-v"><EXP>ser</EXP></SYNSET>
  <SYNSET SCORE="0.00515927770112184" CODE="01666138-v"><EXP>trabajar</EXP></SYNSET>
  <SYNSET SCORE="0.0420541124242606" CODE="01775163-v"><EXP>existir</EXP><EXP>haber</EXP
      ></SYNSET>
  <SYNSET SCORE="0.00347951286819845" CODE="01840295-v"><EXP>caracterizar</EXP><EXP>
      personificar</EXP><EXP>encarnar</EXP></SYNSET>
  <SYNSET SCORE="0.0540524326594277" CODE="01811792-v"><EXP>estar</EXP><EXP>haber</EXP><
      /SYNSET>
  <SYNSET SCORE="0" CODE="01843641-v"><EXP>costar</EXP><EXP>valer</EXP></SYNSET>
  <SYNSET SCORE="0.000119983202351671" CODE="01552250-v"/>
  <SYNSET SCORE="0.0418741376207331" CODE="01781222-v"><EXP>acaecer</EXP><EXP>suceder</
      EXP><EXP>ocurrir</EXP></SYNSET>
  <SYNSET SCORE="5.99916011758354e-05" CODE="01782836-v"><EXP>durar</EXP><EXP>vivir</EXP
      ><EXP>existir</EXP></SYNSET>
  <SYNSET SCORE="0.0161977323174756" CODE="01817610-v"><EXP>equivaler</EXP><EXP>
      significar</EXP><EXP>ser equivalente</EXP></SYNSET>
</TERM>
<TERM ID="LA081794-0225-19" LEMA="THE" POS="DT">
  <WF>THE</WF>
</TERM>
<TERM ID="LA081794-0225-20" LEMA="first" POS="NNP">
  <WF>FIRST</WF>
  <SYNSET SCORE="1" CODE="09974999-n"><EXP>primero</EXP><EXP>primera</EXP></SYNSET>
  <SYNSET SCORE="0" CODE="09770029-n"><EXP>primero</EXP><EXP>primera</EXP></SYNSET>
  <SYNSET SCORE="0" CODE="00464759-n"><EXP>primera base</EXP></SYNSET>
  <SYNSET SCORE="0" CODE="10965545-n"><EXP>comienzo</EXP><EXP>umbral</EXP><EXP>principio
      </EXP><EXP>inicio</EXP></SYNSET>
  <SYNSET SCORE="0" CODE="02688619-n"><EXP>primera</EXP></SYNSET>
  <SYNSET SCORE="0" CODE="05023676-n"/>
</TERM>
<TERM ID="LA081794-0225-21" LEMA="TO" POS="TO">
  <WF>TO</WF>
</TERM>
<TERM ID="LA081794-0225-22" LEMA="state" POS="NNP">
  <WF>STATE</WF>
  <SYNSET SCORE="0" CODE="06060831-n"><EXP>departamento de estado</EXP></SYNSET>
  <SYNSET SCORE="0" CODE="06374245-n"><EXP>estado federal</EXP><EXP>estado</EXP></SYNSET
      >
  <SYNSET SCORE="0.376152427781192" CODE="06074189-n"><EXP>estado</EXP><EXP>pais</EXP><
      EXP>republica</EXP><EXP>nacion</EXP></SYNSET>
  <SYNSET SCORE="0.451751690227412" CODE="06079469-n"/>
  <SYNSET SCORE="0.172095881991396" CODE="00016185-n"><EXP>estado</EXP></SYNSET>
  <SYNSET SCORE="0" CODE="10077290-n"/>
  <SYNSET SCORE="0" CODE="06299747-n"><EXP>tierra</EXP><EXP>estado</EXP><EXP>pais</EXP><
      EXP>nacion</EXP></SYNSET>
  <SYNSET SCORE="0" CODE="10386919-n"><EXP>estado de la materia</EXP><EXP>estado fisico<
      /EXP></SYNSET>
</TERM>
<TERM ID="LA081794-0225-23" LEMA="THAT" POS="WDT">
  <WF>THAT</WF>
</TERM>
<TERM ID="LA081794-0225-24" LEMA="harm" POS="NNP">
  <WF>HARM</WF>
  <SYNSET SCORE="0.25" CODE="05522654-n"><EXP>daÃ±o</EXP><EXP>perjuicio</EXP></SYNSET>
  <SYNSET SCORE="0.625" CODE="10257548-n"><EXP>trauma</EXP><EXP>daÃ±o</EXP><EXP>
      contusion</EXP><EXP>herida</EXP><EXP>lesion</EXP><EXP>traumatismo</EXP></SYNSET>
```

```
      <SYNSET SCORE="0.125" CODE="00258668-n"><EXP>daÃšo</EXP><EXP>perjuicio</EXP><EXP>mal</
          EXP><EXP>destrozo</EXP></SYNSET>
    </TERM>
    <TERM ID="LA081794-0225-25" LEMA="COULD" POS="NNP">
      <WF>COULD</WF>
    </TERM>
    <TERM ID="LA081794-0225-26" LEMA="COME" POS="NNP">
      <WF>COME</WF>
    </TERM>
    <TERM ID="LA081794-0225-27" LEMA="FROM" POS="NNP">
      <WF>FROM</WF>
    </TERM>
    <TERM ID="LA081794-0225-28" LEMA="keyboard" POS="NNP">
      <WF>KEYBOARD</WF>
      <SYNSET SCORE="1" CODE="02887166-n"><EXP>teclado</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="02887056-n"/>
    </TERM>
    <TERM ID="LA081794-0225-29" LEMA="misuse" POS="NNP">
      <WF>MISUSE</WF>
      <SYNSET SCORE="1" CODE="00606110-n"><EXP>abuso</EXP><EXP>desaprovechamiento</EXP></
          SYNSET>
    </TERM>
    <TERM ID="LA081794-0225-30" LEMA="OR" POS="NNP">
      <WF>OR</WF>
    </TERM>
    <TERM ID="LA081794-0225-31" LEMA="TOO" POS="NNP">
      <WF>TOO</WF>
    </TERM>
    <TERM ID="LA081794-0225-32" LEMA="much" POS="NNP">
      <WF>MUCH</WF>
      <SYNSET SCORE="1" CODE="09921720-n"><EXP>mucho</EXP></SYNSET>
    </TERM>
    <TERM ID="LA081794-0225-33" LEMA="type" POS="VBG">
      <WF>TYPING</WF>
      <SYNSET SCORE="0.333333333333333" CODE="00418418-v"/>
      <SYNSET SCORE="0.666666666666667" CODE="00679313-v"><EXP>mecanografiar</EXP></SYNSET>
    </TERM>
    <TERM ID="LA081794-0225-34" LEMA="." POS=".">
      <WF>.</WF>
    </TERM>
  </HEADLINE>
  <TEXT>
  ...
  </TEXT>
</DOC>
```

Itzulpen onena gaztelaniara

```
<DOC>
  <DOCNO>LA081794-0225</DOCNO>
  <DOCID>LA081794-0225</DOCID>
  <HEADLINE>
    <TERM ID="LA081794-0225-1" LEMA="2" POS="CD">
      <WF>2</WF>
    </TERM>
    <TERM ID="LA081794-0225-2" LEMA="firm" POS="NNPS">
      <WF>FIRMS</WF>
      <SYNSET SCORE="1" CODE="06007316-n"/>
    </TERM>
    <TERM ID="LA081794-0225-3" LEMA="ADOPT" POS="NNP">
      <WF>ADOPT</WF>
    </TERM>
    <TERM ID="LA081794-0225-4" LEMA="label" POS="NNP">
      <WF>LABELS</WF>
      <SYNSET SCORE="0.428571428571429" CODE="05381676-n"><EXP>etiqueta</EXP></SYNSET>
      <SYNSET SCORE="0.142857142857143" CODE="05427590-n"/>
      <SYNSET SCORE="0.285714285714286" CODE="05128468-n"/>
      <SYNSET SCORE="0.142857142857143" CODE="10472538-n"/>
    </TERM>
    <TERM ID="LA081794-0225-5" LEMA="warn" POS="VBG">
      <WF>WARNING</WF>
      <SYNSET SCORE="1" CODE="00589833-v"><EXP>avisar</EXP><EXP>notificar</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="00590070-v"/>
```

```
</TERM>
<TERM ID="LA081794-0225-6" LEMA="computer" POS="NNP">
  <WF>COMPUTER</WF>
  <SYNSET SCORE="0" CODE="07135102-n"/>
  <SYNSET SCORE="1" CODE="02481557-n"><EXP>ordenador</EXP><EXP>procesador</EXP></SYNSET>
</TERM>
<TERM ID="LA081794-0225-7" LEMA="user" POS="NNPS">
  <WF>USERS</WF>
  <SYNSET SCORE="0.25" CODE="07252689-n"/>
  <SYNSET SCORE="0" CODE="07229219-n"/>
  <SYNSET SCORE="0.75" CODE="07658488-n"><EXP>usuario</EXP></SYNSET>
</TERM>
<TERM ID="LA081794-0225-8" LEMA="ABOUT" POS="NNP">
  <WF>ABOUT</WF>
</TERM>
<TERM ID="LA081794-0225-9" LEMA="danger" POS="NNP">
  <WF>DANGER</WF>
  <SYNSET SCORE="0.585125998770744" CODE="00512734-n"><EXP>peligro</EXP><EXP>riesgo</EXP
      ></SYNSET>
  <SYNSET SCORE="0" CODE="06365821-n"/>
  <SYNSET SCORE="0.221266133988937" CODE="10427605-n"/>
  <SYNSET SCORE="0.19360786724032" CODE="10427326-n"/>
</TERM>
<TERM ID="LA081794-0225-10" LEMA="OF" POS="IN">
  <WF>OF</WF>
</TERM>
<TERM ID="LA081794-0225-11" LEMA="injury" POS="NN">
  <WF>INJURY</WF>
  <SYNSET SCORE="0.569760295021511" CODE="10257548-n"><EXP>trauma</EXP><EXP>dano</EXP><
      EXP>contusion</EXP><EXP>herida</EXP><EXP>lesion</EXP><EXP>traumatismo</EXP></
      SYNSET>
  <SYNSET SCORE="0.430239704978489" CODE="05451786-n"/>
  <SYNSET SCORE="0" CODE="00479770-n"/>
  <SYNSET SCORE="0" CODE="05469280-n"/>
</TERM>
<TERM ID="LA081794-0225-12" LEMA=";" POS=":">
  <WF>;</WF>
</TERM>
<TERM ID="LA081794-0225-13" LEMA="technology" POS="NNP">
  <WF>TECHNOLOGY</WF>
  <SYNSET SCORE="0.827904118008605" CODE="00607693-n"><EXP>ingenieria</EXP><EXP>
      tecnologia</EXP></SYNSET>
  <SYNSET SCORE="0.172095881991395" CODE="04660658-n"/>
</TERM>
<TERM ID="LA081794-0225-14" LEMA=":" POS=":">
  <WF>:</WF>
</TERM>
<TERM ID="LA081794-0225-15" LEMA="COMPAQ" POS="NNP">
  <WF>COMPAQ</WF>
</TERM>
<TERM ID="LA081794-0225-16" LEMA="," POS=",">
  <WF>,</WF>
</TERM>
<TERM ID="LA081794-0225-17" LEMA="MICROSOFT" POS="NNP">
  <WF>MICROSOFT</WF>
</TERM>
<TERM ID="LA081794-0225-18" LEMA="be" POS="VBP">
  <WF>ARE</WF>
  <SYNSET SCORE="0.0113384126222329" CODE="01787769-v"/>
  <SYNSET SCORE="0.181174635551023" CODE="01784339-v"/>
  <SYNSET SCORE="0.644489771431999" CODE="01775973-v"><EXP>ser</EXP></SYNSET>
  <SYNSET SCORE="0.00515927770112184" CODE="01666138-v"/>
  <SYNSET SCORE="0.0420541124242606" CODE="01775163-v"/>
  <SYNSET SCORE="0.00347951286819845" CODE="01840295-v"/>
  <SYNSET SCORE="0.0540524326594277" CODE="01811792-v"/>
  <SYNSET SCORE="0" CODE="01843641-v"/>
  <SYNSET SCORE="0.000119983202351671" CODE="01552250-v"/>
  <SYNSET SCORE="0.0418741376207331" CODE="01781222-v"/>
  <SYNSET SCORE="5.99916011758354e-05" CODE="01782836-v"/>
  <SYNSET SCORE="0.0161977323174756" CODE="01817610-v"/>
</TERM>
<TERM ID="LA081794-0225-19" LEMA="THE" POS="DT">
  <WF>THE</WF>
</TERM>
<TERM ID="LA081794-0225-20" LEMA="first" POS="NNP">
  <WF>FIRST</WF>
  <SYNSET SCORE="1" CODE="09974999-n"><EXP>primero</EXP><EXP>primera</EXP></SYNSET>
```

```
      <SYNSET SCORE="0" CODE="09770029-n"/>
      <SYNSET SCORE="0" CODE="00464759-n"/>
      <SYNSET SCORE="0" CODE="10965545-n"/>
      <SYNSET SCORE="0" CODE="02688619-n"/>
      <SYNSET SCORE="0" CODE="05023676-n"/>
    </TERM>
    <TERM ID="LA081794-0225-21" LEMA="TO" POS="TO">
      <WF>TO</WF>
    </TERM>
    <TERM ID="LA081794-0225-22" LEMA="state" POS="NNP">
      <WF>STATE</WF>
      <SYNSET SCORE="0" CODE="06060831-n"/>
      <SYNSET SCORE="0" CODE="06374245-n"/>
      <SYNSET SCORE="0.376152427781192" CODE="06074189-n"/>
      <SYNSET SCORE="0.451751690227412" CODE="06079469-n"/>
      <SYNSET SCORE="0.172095881991396" CODE="00016185-n"/>
      <SYNSET SCORE="0" CODE="10077290-n"/>
      <SYNSET SCORE="0" CODE="06299747-n"/>
      <SYNSET SCORE="0" CODE="10386919-n"/>
    </TERM>
    <TERM ID="LA081794-0225-23" LEMA="THAT" POS="WDT">
      <WF>THAT</WF>
    </TERM>
    <TERM ID="LA081794-0225-24" LEMA="harm" POS="NNP">
      <WF>HARM</WF>
      <SYNSET SCORE="0.25" CODE="05522654-n"/>
      <SYNSET SCORE="0.625" CODE="10257548-n"><EXP>trauma</EXP><EXP>daÃ§o</EXP><EXP>
          contusion</EXP><EXP>herida</EXP><EXP>lesion</EXP><EXP>traumatismo</EXP></SYNSET>
      <SYNSET SCORE="0.125" CODE="00258668-n"/>
    </TERM>
    <TERM ID="LA081794-0225-25" LEMA="COULD" POS="NNP">
      <WF>COULD</WF>
    </TERM>
    <TERM ID="LA081794-0225-26" LEMA="COME" POS="NNP">
      <WF>COME</WF>
    </TERM>
    <TERM ID="LA081794-0225-27" LEMA="FROM" POS="NNP">
      <WF>FROM</WF>
    </TERM>
    <TERM ID="LA081794-0225-28" LEMA="keyboard" POS="NNP">
      <WF>KEYBOARD</WF>
      <SYNSET SCORE="1" CODE="02887166-n"><EXP>teclado</EXP></SYNSET>
      <SYNSET SCORE="0" CODE="02887056-n"/>
    </TERM>
    <TERM ID="LA081794-0225-29" LEMA="misuse" POS="NNP">
      <WF>MISUSE</WF>
      <SYNSET SCORE="1" CODE="00606110-n"><EXP>abuso</EXP><EXP>desaprovechamiento</EXP></
          SYNSET>
    </TERM>
    <TERM ID="LA081794-0225-30" LEMA="OR" POS="NNP">
      <WF>OR</WF>
    </TERM>
    <TERM ID="LA081794-0225-31" LEMA="TOO" POS="NNP">
      <WF>TOO</WF>
    </TERM>
    <TERM ID="LA081794-0225-32" LEMA="much" POS="NNP">
      <WF>MUCH</WF>
      <SYNSET SCORE="1" CODE="09921720-n"><EXP>mucho</EXP></SYNSET>
    </TERM>
    <TERM ID="LA081794-0225-33" LEMA="type" POS="VBG">
      <WF>TYPING</WF>
      <SYNSET SCORE="0.333333333333333" CODE="00418418-v"/>
      <SYNSET SCORE="0.666666666666667" CODE="00679313-v"><EXP>mecanografiar</EXP></SYNSET>
    </TERM>
    <TERM ID="LA081794-0225-34" LEMA="." POS=".">
      <WF>.</WF>
    </TERM>
  </HEADLINE>
  <TEXT>
  ...
  </TEXT>
</DOC>
```