V

Tresna linguistikoak informazioa atzitzeko

Eneko Agirre eta Iñaki Alegria
Ixa Taldea. Euskal Herriko Unibertsitateko irakasleak
Grupo Ixa. Profesores de la Universidad del País Vasco

Laburpena: Hemeroteka digitalen informazio-bolumena izugarri hazi da azken urteetan. Ondorioz, bilaketak zaildu egin dira, besteak beste eleaniztasunagatik, erabiltzaileak askotan dokumentu eleanitzen artean bilatu behar duelako. Tresna linguistikoek laguntzen dute informazioaren atzipen erosoago eta zehatzagoa lortzeko, baina beraien erabilera ez dago guztiz zabaldua. Bilaketa-teknologia, eta bereziki tresna linguistikoek egiten duten ekarpena azalduko dute egileek artikulu honetan. Bilaketaren inguruko problematika aztertu ondoren, hemeroteka digitalen gaineko bilaketez arituko dira. Horrez gain, tresna linguistikoek egin dezaketen ekarpena eta merkatuan dauden hainbeste aplikazio azalduko dituzte.

Resumen: El volumen de información que contienen las hemerotecas digitales ha aumentado considerablemente en los últimos años. Como consecuencia, las búsquedas se hacen más difíciles, entre otras cosas debido al multilingüismo, ya que el usuario en muchas ocasiones debe buscar entre documentos en varias lenguas. Las herramientas lingüísticas facilitan un resultado más cómodo y preciso, pero su utilización no está muy extendida. En este artículo, los autores explican la tecnología de búsqueda, más concretamente las aportaciones de las herramientas lingüísticas. Tras analizar la problemática acerca de las búsquedas, nos hablan sobre las búsquedas en hemerotecas digitales. Por último, nos explican qué aportaciones pueden hacer las herramientas lingüísticas, además de otras aplicaciones que existen en el mercado.

1. SARRERA

Hemeroteka digitalek arrakasta handia izan dute azken urteetan; horren ondorioz bertan metatzen den informazio-bolumena izugarri hazi da eta bilaketak zaildu egin dira. Horren aurrean bilaketen teknologia asko garatu da, informazioaren atzipen erosoa eta zehatza helburu. Gaur egun bilaketen teknologia heldua da, *Google* eta *Yahoo* bezalako bilatzaileen garapenaren eraginez. Teknologia hori hemeroteketan aplikatzen denean emaitzak nahiko onak badira ere, hobekuntzak eta doikuntzak egin behar dira hemeroteken ezaugarrien arabera, batez ere metadatuen trataera egokia lortzeko. Bilaketa korapilatzen duen ezaugarri bat eleaniztasuna da: askotan komenigarria edo ezinbestekoa

da dokumentu eleanitzen artean bilatzea edo galderak hizkuntza desberdinetan onartzea.

Tresna linguistikoek laguntzen dute informazioaren atzipen erosoago eta zehatzagoa lortzeko, baina beraien erabilera ez dago guztiz zabaldua, beharbada oraindik nahiko garatuta ez daudelako edo bilaketa-sistemari eransten dioten konplexutasunagatik. Bilaketa-teknologia, eta bereziki tresna linguistikoek egiten duten ekarpena azalduko dugu artikulu honetan. Bilaketaren inguruko problematika aztertu ondoren, hemeroteka digitalen gaineko bilaketez arituko gara. Horrez gain, tresna linguistikoek egin dezaketen ekarpena eta merkatuan dauden hainbeste aplikazio azalduko ditugu.

2. INFORMAZIOAREN BILAKETA ETA BERE ALDAERAK

Informazioaren berreskurapena (*Information Retrieval*, IR) ohiko arlo bat izan da informatikaren garapenaren hasieratik. Konputagailuek informazio kopuru handiak biltegiratzea posible egiten dutenez, informazio hori modu zehatz, eroso eta eraginkorrean berreskuratzea beti izan da aztergai garrantzitsua. Informazioaren berreskurapenaren kontzeptua testu-masa handien biltegiratze/berreskuratzearekin lotu ohi da informatikaren munduan [2]. Datubase dokumentalak izan dira arlo honetako aplikazio garrantzitsuenak, eta bertan lantzen dira gaian gakoa diren bi urratsak: dokumentuen indexazioa eta ondorengo bilaketa.

Internet fenomenoak bultzatu egin du arlo honen garapena, testu digitalak izugarri ugaldu direlako. IRren ohiko aplikazioez gain (testu legalak, medikuntzakoak, hemerotekak, dokumentazio-zentroak, ...) Internet/Intranet eremuko aplikazio garrantzitsuenak kokatzen dira arlo honetan: *Google* moduko bilatzaileak eta *Yahoo* moduko direktorioak.

Duela gutxi arte, tresnen abiadura motela dela-eta, hizkuntza-ingeniaritzak ez du oso paper garrantzitsua jokatu arlo honen garapenean. Dena den, tresna linguistikoak hobetu diren heinean eta dokumentu digitalen eleaniztasuna areagotzearekin batera, tresna linguistikoen erabilpena garrantzia hartzen joan da.

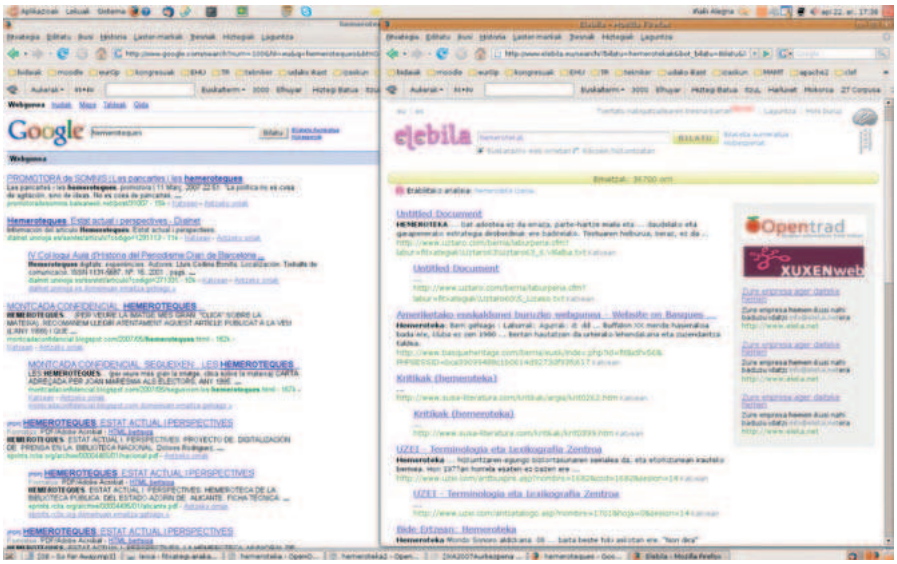
Ixa taldean[1] arlo honetan lan sakona egin dugu azken urteetan, ikuspegi eleanitz batetik, baina euskararen gaineko bilaketen problematikari erreparatuz.

3. OHIKO BILAKETA

Bilaketen teknologia azken urteetako teknologia arrakastatsuena izan da. Interneteko zabalkundearekin batera bilatzaileak eguneroko tresnak dira jende

[1] http://ixa.si.ehu.es

asko eta askorentzat. *Google*[2] eta neurri txikiagoan *Yahoo*[3] eta *LiveSearch*[4] dira bilatzaile osoen eta erabilienak. Hizkuntza nagusiak baino ez dituzte kontuan hartzen, eta beraz, euskara bezalako hizkuntza txikientzat arazoak daude ohiko bilatzaileentzat. Horren aurrean bilaketako gaia desitxuratu (*energia*-ri buruz diharduten euskarazko dokumentuak bilatzeko *energia eta du* gakoak erabiltzea adib.) edo hizkuntza horretarako propio egindako bilatzaile bat erabiltzea (euskarako Elebila[5]) [4] dira aukerak. 1. irudian bilatzaile hauen adibideak ikus daitezke.



1. irudia

Google-en eta Elebilaren interfazeak

Sistema hauen teknologia aski ezaguna da duela urte batzuetatik hona. Aurkitutako dokumentuak ordenatzeko garaian, dokumentuak duen edukiaz gain hainbat faktore hartzen dira kontuan: dokumentuak nondik erreferentziatzen dituzten, zenbat aldiz eta erreferentzia egiten duen gunearen esanguratasuna (*PageRank* algoritmoa [3] izan zen *Google*-en arrakastaren hasierako gakoa). Informazio hori hipertesteken topologiaren azterketaren bidez lortzen da. Hala ere, azken urteetan bilatzaileen arteko lehiakortasuna dela-eta, bila-

---

2 www.google.com

3 www.yahoo.com

4 www.live.com

5 www.elebila.com

tzaileen barne-funzionamenduari buruzko kaleratzen den informazioa murritzagoa da. Gainera, bilatzaileek webguneetara iristeko bide nagusi bihurtu diren heinean, enpresak saiatzen dira algoritmo horiei iskin egiten lehen postutan agertzeko (honi *search spam* deitzen zaio) eta bilatzaileek publiko ez diren neurriak hartzen dituzte horren aurka.

Teknologia honen osagai nagusiak hiru dira: robota, sarea miatzeko dokumentuen bila; indexatzailea dokumentuetatik indizeak gehitzeko sistema; eta bilatzaile bera, galdera baten aurrean indizeak kontsultatu eta emaitzak (eta dokumentuen estekak) itzultzen dituena. Informazio kopurua erraldoia da, eta horrek hardware eta komunikazioen aldetik konplexutasuna handitzen du izugarri.

Internet osoaren eskalan lan egin beharrean Intranetean aritzean azken arazo hori desagertzen da, baina teknologia antzekoa da. Dena den ondoren aztertuko dugu gaia, hemerotekez eta liburutegi digitalez arituko garenean.

Emaitzak ebaluatzerakoan bi neurri erabiltzen dira doitasuna (*precision*) eta estaldura (*recall*) [2]. Lehena eskuratutako dokumentuen esanguratasuna neurtzen du, hau da, eskuratutako dokumentuen artean zenbat dira esanguratsu edo interesgarri bilaketa-beharrerako. Estaldura, berriz, sistemaren eraginkortasuna neurtzen du, hau da, zeuden dokumentu esanguratsuen artean zenbat itzuli diren. Doitasuna kalkulatzea erraza da, dokumentalista batek eskuratutako dokumentuak aztertu eta esanguratsuak zeintzuk diren markatuta. Estaldura neurtzea oso zaila da, bildumaren dokumentu guztiak begiratu beharko liratekeelako, eta orokorrean modu erlatiboan neurtzen da, galdera beraren aurrean sistema batek beste batek baino dokumentu esanguratsu gehiago edo gutxiago itzultzen duenetz egiaztatuz.

Bilaketa-sistemetan gertatzen diren egoera korapilatsuak honako hauek dira: galdera baten aurrean dokumenturik ez aurkitzea, edo erantzun gehiegi agertzea. Lehena konpontzeko estaldura handitu behar da, semantika erabiliz adibidez (ikusi geroago). Bigarrenari aurre egiteko emaitzak ordena egokian aurkeztea da gakoa, baina galderak fintzen laguntzeko tresnak ere badaude.

Bilaketaren arloan azken urteetan aurrerakuntza nabarmenak egon dira. Gure gaiarekin lotuta honako hauek azpimarra daitezke:

— CLIR (*Cross-Lingual IR*): bilaketa eleanitzean galderak eta dokumentuak hainbat hizkuntzatan egon daitezke eta bilatzailea gai izan behar da hizkuntza desberdinetako dokumentuak erlazionatzeko [6]. Hiztegi bat nahikoa izan daiteke horretarako, baina itzulpen automatikoa gero eta gehiago erabiltzen da.
— QA (*Question Answering*): dokumentuak bilatu beharrean erantzun zehatzak lortu nahi dira sistema hauetan (nork, non, zergatik...). Helburu horrekin dokumentuen prozesaketa sakonagoa behar da, hizkuntza-teknologiak ezinbesteko tresna izanik. Emaitzak oraindik ez dira oso ikusgarriak baina ikerketa handia egiten da arlo honetan.
— Bilaketa multimodala: digitalizazioa dela-eta, dokumentuak bilatzeaz gain, argazkiak, irratiko edo telebista/bideoko programak bila daitezke.

Bilaketa-sistema hauek metadatuetan (fitxa dokumentala) oinarri daitezke edo bestelako tekniketan (inguruko testua argazkietarako, ahots-testu bihurketa, etab.); eta askotan metodo konbinatuak dira egokienak. Gai hau artikulu honen esparrutik kanpo geratzen da.

4. BILAKETA HEMEROTEKETAN ETA LIBURUTEGI DIGITALETAN

Interneteko bilatzaile bat eta hemeroteka batean edo liburutegi digital batean behar den bilatzailearen artean desberdintasun nagusiak honako hauek dira:

— Dimentsioa: hemerotekak (eta liburutegi digitalak) handiak izan ohi dira, baina ez Internet sare osoa bezain handiak. Informazioa lokala izan ohi da, eta horren ondorioz, gehienetan, ez da aipatu den robota (sarea usnatzen duen elementua) behar.
— Metainformazioa: sarean deskribatzaileak difusoak dira, eta dokumentuen arteko estekak eta izenburuak dira emaitzak ordenatzeko erabiltzen diren oinarrizko irizpideak. Hemeroteketan metadatuak egon ohi dira, osagai bakoitza fitxa dokumental batez deskribatzen baita. Beraz, bilatzailea metadatu horiei etekina ateratzen saiatu behar da. Bestalde dokumentuen arteko aipamenak egon ohi dira, baina gehienetan ez daude modu esplizituan (ez daude estekak), baina oso gai interesgarria da bilaketaren doitasuna handitzeko.

Metadatuetan ematen den aniztasun eta elkar ulertzeko zailtasunak direla eta, gomendagarria da estandarrak jarraitzea. MARC eta *Dublin Core* dira proposamen interesgarrienak.

Aurrekoa kontuan hartuta sistema hauetako bilaketa metadatuetan oinarritu ohi da, baina estaldura galdu nahi ez bada, testua ere erabil beharko da. Kontuan hartu behar da sistema hauetan estaldura oso garrantzitsua dela, Internet-en ez bezala, hemen dokumentuen erredundantzia txikia da eta, ondorioz, bilaketa askotarako dokumentu urri batzuk besterik ez daude. Gainera, erabiltzaileak askotan profesionalak dira eta bilaketa zehatzagoak eskatzen dituzte. Bestalde, ohiko bilatzaileen eraginez gero eta gutxiago erabiltzen dira bilatzeko formulario konplexuak, beraz, gomendagarria da metadatuen integrazioa bilaketa sinplean.

Teknologiari dagokionez, badaude hainbat produktu interesgarri bilaketak egiteko mota honetako eszenatokietan. *Autonomy*[6] bilatzailea liderra omen da enpresa handietan, baina bere sistema oso garestia eta itxia da. Horren ordez *Lucene*[7] erabil daiteke, irekia, moldagarria eta librea baita. Liburutegi digital osoak eraikitzeko, bilaketa eta guzti, beste programa libre interesgarri bat

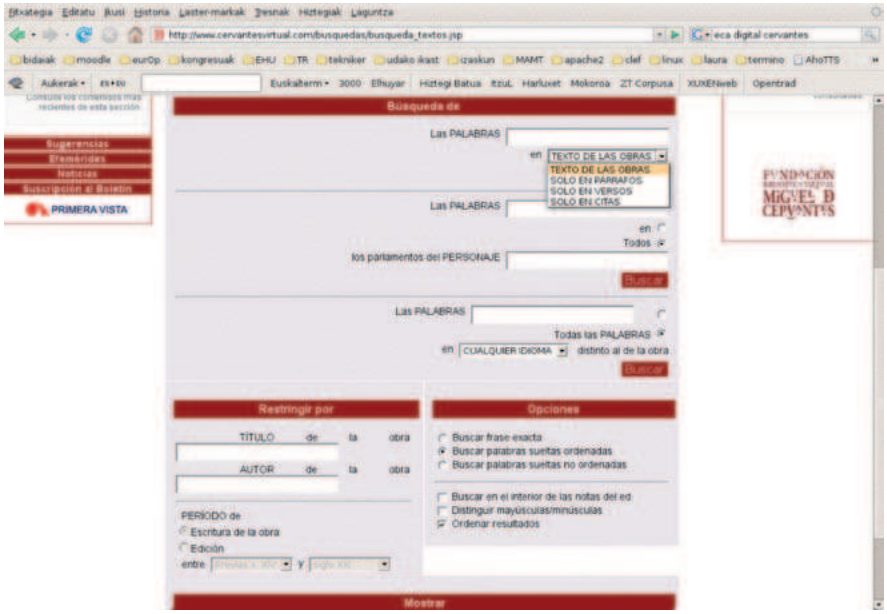
---

[6] www.autonomy.com

[7] http://lucene.apache.org

dago, *Greenstone*[8]. Dena den, oraindik ohikoak dira neurrira egindako soluzio itxiak.

2. irudian mota honetako bilatzaile baten adibidea dugu. Cervantes liburutegi digitalean bilatzeko interfazea[9] ikus daiteke bertan. Hainbat metadaturen arabera bilatzeko aukera eskaintzen du.



2. irudia

Bilaketa liburutegi digital batean

5. TRESNA LINGUISTIKOAK BILAKETARAKO APLIKAZIOAK

Hizkuntza-ingeniaritza arloko tresna linguistikoek bilatzaileen doitasuna edota estaldura handitzen lagundu dezakete, batez ere hemeroteketako zein liburutegi digitaletako bilatzaileetan. Konplexutasunaren arabera aurkeztuko ditugu, atal bakoitzean oinarrizko teknologia eta aplikazioa azalduz.

---

[8] www.greenstone.org

[9] www.cervantesvirtual.com

6. MORFOLOGIAN OINARRITUTAKO TRESNAK

Oinarrizko teknologia hitza-lema arteko erlazioa bilatzea da. Analisia esaten zaio hitzetik lema edo forma kanonikoa (hiztegiko sarreraren baliokidea) lortzen den eragiketari, eta sorkuntza lematik forma posible guztiak lortzen denari. Flexio handiko hizkuntzatan tratamendu morfologikoa ezinbestekoa da estaldura oso txikia ez izateko.

Bilatzaileetan prozesaketa morfologikoa bi modutan egin daiteke. Lehenengoan indexatze-prozesuan hitzen ordez, edo hitzekin batera, lemak gorde egiten dira indizeetan. Bigarrenean hitzak gorde egiten dira, baina bilaketa egitean sorkuntza egiten da, sortutako hitz guztiak bilatuz, *edo* eragileaz konbinaturik.

Teknologia linguistikoa izan daiteke (lemak, aurrizkiak, atzizkiak, paradigmak, aldaketa fonologikoak etab. erabiliz), edo hurbilpen batez ebatzi, azken kasu horretan *stemming* edo sasilematizazioa esaten zaio eragiketari. Bilatzaile handietan sasilematizazio sinplea erabiltzen da, baina euskararen kasuan arriskutsua izan daiteke doitasun-galera handi sor daitekeelako.

3. irudia

Galdera zuzentzeko proposamenak

Aipatutako *elebila* bilatzaileak (1. irudia) integratuta du morfologia sorkuntzaren bidez [4]. Beraz, hitz bat ematen dugunean bilatzeko bere lemaren forma guztiak (edo garrantzitsuenak) sortzen dira estaldura egokia ziurtatzeko.

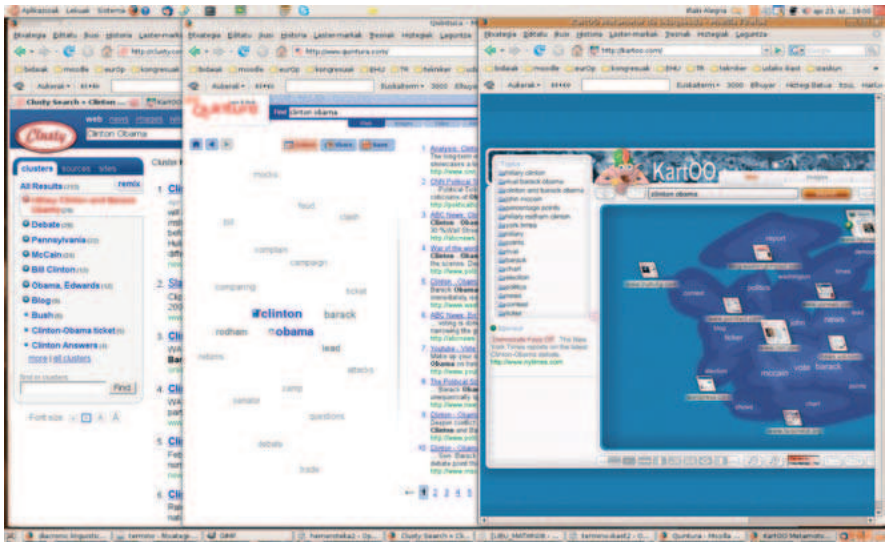
Horrez gain, morfologia eta estatistika konbina daitezke galderetan erroreak detektatzeko eta proposamen egokiak emateko, 3. irudian ikus daitekeen moduan.

7. SINTAXIAN OINARRITUTAKO APLIKAZIOAK

Testuen analisi sintaktiko sakona konputagailuz egitea ikergaia da gaur egun. Dena den analisi partzialak lortzen dituzten programak oso hedatuta dau-

de eta ondo funtzionatzen dute. Orokorrean IR arloan izen sintagmak bilatzen dira, informazio garrantzitsua eskaintzen dutelako: termino berriak, pertsonak, enpresak/erakundeak, tokiak...

Oinarrizko elementu horiek eta estatistikak konbinatuz aplikazio interesgarriak sor daitezke, dokumentuak estekatuz edo multzokatuz adibidez. Gainera erantzun gehiegi itzultzen dituzten galderak fintzen lagun dezakete aplikazio horiek.



4. irudia

Sintaxian eta estatistikan oinarritutako sistemen adibideak

4. irudian agertzen diren webguneak horren adibideak dira: www.clusty.com, www.quintura.com, www.kartoo.com Ikus daitekeenez, batzuetan emaitzak erabiltzen dira bistaratze bereziak egiteko (multzoak, erlazioak...).

Euskarazko guneen artean www.zientzia.net aipa dezakegu, lematizazioa eta multzokatzea integratzen ditu-eta.

8. ELEANIZTASUNA. CLIR

Aplikazio eleanitzak egiteko hainbat aukera daude: hiztegi elebidun digitalak, ontologia eleanitzak eta itzulpen automatikoa. Ontologia eleanitzaren adibide gisa hurrengo atalean azalduko den *WordNet* dugu. Itzulpen automatiko zehatza [6] aspaldiko amets betegabea da, baina azken urteetan aurrerapen handiak egin dira, batez IR sistemetan integra daitezken sistema arin eta azkarren aldetik. Doitasuna ez da oso handia, baina aplikazioaren helburua itzultzea

ez denez kalitate onargarria izan daiteke IR aplikazioetarako. Antzekotasun handiko hizkuntzen artean itzultzeko edo baliabide asko duten hizkuntza nagusien artean itzultzeko, gaur programa nahiko zehatzak daude.

Euskararen kasuan, oraindik gauza asko egin behar dira baina gaztelaniatik euskara itzultzen duen lehen sistema dago eskuragarri[10]. Kalitatea oraindik ez da egokia itzulpen profesionalean erabiltzeko baina halako aplikazioetan erabil daiteke. 5. irudian adibide bat ikus daiteke.

Eleaniztasunean oinarritutako aplikazioen adibideak hemeroteka eleanitzak ditugu. Hainbat diseinu egin daiteke, bilduma hizkuntza bakar batean egotea, baina galderak edozein hizkuntzetan egin ahal izatea; galderak hizkuntza bakar batean baina dokumentuak eleanitzak izatea, edo aurreko bien konbinazioa., galderak eta dokumentuak eleanitzak. Are gehiago, argazkien edo bideoaren gaineko bilaketa eleanitza egin daiteke.

Horretarako bilatzailea hainbat aukera ditu, galderak edota dokumentuak itzultzea hitzez hitz hiztegien bitartez, edo itzultzea itzulpen automatikoaren bidez. Aukera gehiago daude, ontologia batera proiekta daitezke galderak eta dokumentuak. Kasu horretan bilaketa semantikoaz hitz egin daiteke.

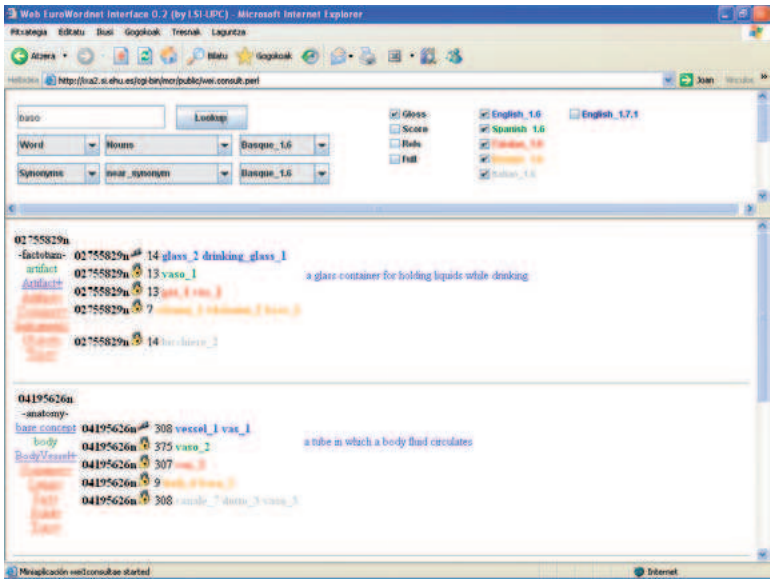
5. irudia

Opentrad itzultzailearen adibidea

10 www.opentrad.com

9. SEMANTIKA. BILAKETA SEMANTIKOA

Bilaketa semantikoaren bidez hitzen bidezko bilaketan dauden arazo biri erantzun nahi zaie: adiera anbiguotasuna batetik (*baso* bilatzen badugu arbolen basoa edo edateko basoari buruzko dokumentuak atzituko genituzke), eta sinonimia edo espezializazioa bestetik (*restaurante* bilatzen badugu, sagardotegi edo erretegi hitzak azaltzen diren dokumentuak ez genituzke lortuko). Bilaketa kontzeptuen bitartez egingo balitz, bi arazo horiek arinduko lirateke.



6. irudia

WordNet interfazearen adibidea

Gainera, kontzeptuen inbentario diren ontologiak eleaniztunak balira, orduan hizkuntzen arteko bilaketak ere hobeto egingo lirateke, kontzeptu gehienak konpartitzen baitira hizkuntzen artean. Web 2.0-n erabiltzen diren folksonomiak ez bezala, analisi semantikorako taxonomia eta ontologia formalagoak erabili ohi dira, eta horien artean WordNet (Ingeleserako ontologia elebakarra) eredua jarraitzen duten wordnet eleaniztunak dira erabilienak, IXA taldean beste erakunde batzuekin elkarlanean landu den *Multilingual Central Repository* (MCR) kasu [1]. MCR delakoak ingelera gaztelera, italiera, katalanera eta euskara biltzen ditu bere baitan. Bertan hizkuntzetatik independente izan nahi duen kontzeptu inbentario bat dago, eta kontzeptu horiei buruzko informazio semantiko aberatsa jasotzen da, taxonomia egiturak barne. Azken urteetan egindako ikerketari esker, kontzeptuei buruzko hainbat informazio

sartu izan da MCRen. Beheko irudian euskarazko baso-ri dagozkion bi adiera azaltzen dira, beste hizkuntzatako itzulpenekin batera, eta kontzeptuaren hainbat tasun semantikorekin batera (adibidez, gizonak egindako dela edateko basoa, edukitzeko ontzi baten propietateak dituela, instrumentu bezala erabili daitekeela, etab.). Hurrengo irudian azaltzen dira ere

10. ARGAZKIEN BILAKETA SEMANTIKO ELEANITZA

Ixa taldean, *Meaning*[11] proiektu europarraren barruan, MCR/WordNet erabili dugu argazkien bilaketa eleanitza hobetzeko [1]. EFEko fototeka erabili zen dokumentu-bildumatzat, eta bilaketa semantiko eleanitzaren bitartez lortzen genuen estaldura igotzea. Eleaniztasuna islatzen da 7. irudian. Ingelesezko galdera baten erantzunez gaztelaniazko fitxa duten argazkiak ere agertzen dira.

Alde semantikoaren onura ebaluatu ahal izateko, pertsonak eta hiriak ez ziren argazkiei buruzko 20 ataza definitu ziren. EFEko hiru dokumentalista aritu ziren 20 ataza horietan eskatzen diren argazkiak bilatu nahian, eta ikusi zen sistema semantikoarekin lehenagoko sistemarekin baino argazki gehiago lortzen zirela (24 berriarekin eta 20 aurrekoarekin) ekintza gutxiagoz (168 klik berriarekin eta 295 klik aurrekoarekin). Alde semantikoarekin lortzen diren emaitzen adibide gisa bi adibide aipatuko ditugu:

— Ingurumenarekin lotutako epaile baten argazkia bilatzean, semantikarik gabeko sisteman fitxa dokumentalean «magistrado» agertzen zireneko argazkiak ez ziren lortzen.
— Kolonbiako biolentziaren inguruko ehorzketa baten argazkia eskatzean ohiko sisteman zaila zitzaien argazkia aurkitzea fitxan «entierro» agertu beharrean «sepelio» agertzen zelako.

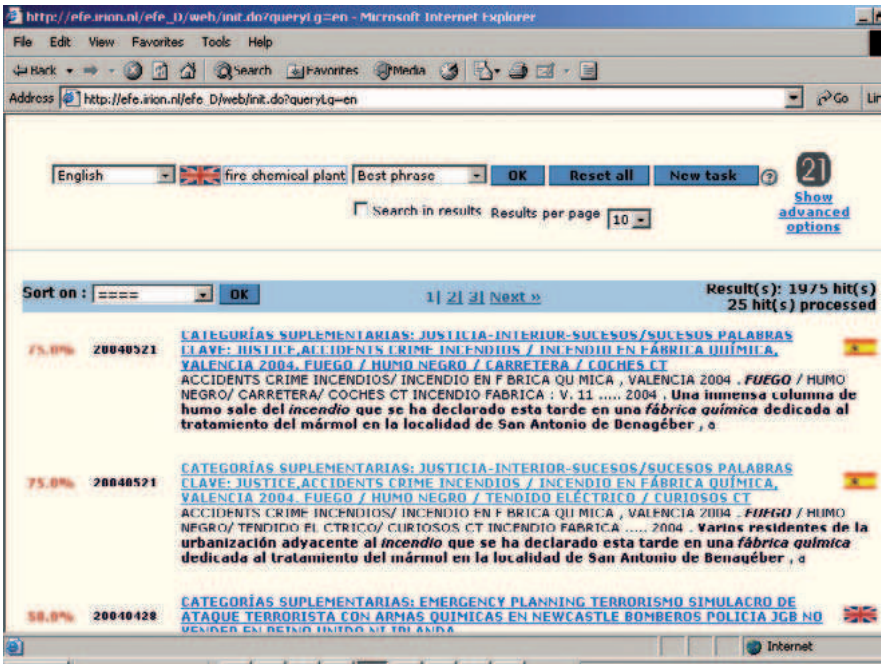
11. ONDORIOAK

Artikuluan azaldu dugunez, teknologia linguistikoak ekarpen handia egin dezake bilaketen arloan, bereziki hemeroteketan eta liburutegi digitaletan. Bilatzaile orokor ezagunenetan oraindik gutxi erabiltzen badira ere, erabilera handitzen doa, orokorrean estaldura handitzeko oso tresna interesgarria direlako. Tresna linguistiko hauek (sintaxia, itzulpena, semantika) ez dira gehiago erabiltzen oraingoz ez direlako nahiko eraginkorrak, prozesaketa-denboraren aldetik, prozesatzeko behar den memoriaren aldetik edo doitasunaren aldetik.

Dena den, gero eta gehiago erabiltzen ari dira, eta aipatutako mugak gainditzen diren neurrian erabilera handiagotu egingo da zalantzarik gabe.

---

[11] www.lsi.upc.edu/~nlp/meaning/



7. irudia

EFE fototekaren gaineko bilaketa eleanitza

BIBLIOGRAFIA

[1] AGIRRE E., ALEGRIA I., RIGAU G., VOSSEN P. 2007. MCR for CLIR. *SEPLN aldizkaria, monografia TIIMM*. vol 38, 3-16. ISSN 1135-5948. http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1174895701/publikoak/pdf

[2] BAEZA-YATES R., RIBEIRO-NETO B. 1999. *Modern Information Retrieval*. ACM Press Series/Addison Wesley, 1999.

[3] BRIN S., PAGE L. 1998. The anatomy of a large-scale hypertextual Web search engine Computer Networks and ISDN Systems, Elsevier.

[4] LETURIA I., GURRUTXAGA A., ARETA N., ALEGRIA I., EZEIZA A. 2007. EusBila, a search service designed for the agglutinative nature of Basque. *SIGIR2007—iNEWS'07 workshop*. ISBN 978-84-690-6978-3. http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1190627800/publikoak/pdf

[5] MAYOR, A. 2007. MATXIN: *Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz*. Doktorego-tesia. Euskal Herriko Unibertsitateko; Donostiako Informatika Fakultatea. http://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak/1196444990/publikoak/Matxin2007.pdf

[6] SARALEGI X., ALEGRIA I. 2007. Similitud entre documentos multilingües de carácter técnico en un entorno Web. *SEPLN aldizkaria*, 2007. Sevilla. ISSN 1135-5948. http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1184248312/publikoak/pdf