Semantikan oinarritutako bilaketak: Kyoto proiektua

Iñaki Alegria eta German Rigau

Euskal Herriko Unibertsitateko irakasleak, IXA Taldea
Profesores de la Universidad del País Vasco, Grupo IXA

Laburpena: Dokumentazioaren kudeaketa digitalean informazioa atzitzeko deskribatzaileak erabiltzeaz gain testu bera erabili ahal izatea oso interesgarria da. Deskribatzaileak beraiek ere askotan testuak dira. Hizkuntza ingeniaritza erabilita hainbat aukera zabaltzen dira datu-baseen informazioa atzitzeko garaian: atzipen eleanitza, multzokatze semantikoa, antzekotasunean oinarritutako atzipena, galdera-erantzun sistemak, informazioa inferitzea... Semantikaren inguruko aukeretan sakonduko dugu, Europako beste unibertsitate eta enpresekin Kyoto proiektuan lantzen ari garen ikerketa-ildoak azalduz.

Resumen: Investigaciones basadas en la semántica: Proyecto de Kyoto. En la gestión digital de la documentación puede ser muy interesante usar el propio texto además de los descritores. Muchos descriptores también son texto. Usando técnicas de ingeniería lingüística se abren nuevas opciones de acceso a la información de estas bases de datos: acceso multilingüe, agrupación semántica, acceso basado en similitud, sistemas de pregunta-respuesta, inferencia de información... Profundizaremos en las posibilidades basadas en semántica, exponiendo las lineas de investigación que estamos desarrollando en el proyecto europeo Kyoto.

SARRERA

Eduki digitalen hedapenarekin eskura dugun 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, *Google* eta *Yahoo* bezalako bilatzaileen garapenaren eraginez, bilaketen teknologia heldua da informazio-behar arruntenetarako, dokumentu edo paragrafo batez asetzen diren beharretarako hain zuzen ere. Beste behar batzuetarako berriz gaur egungo teknologia ez da aski, galdera zehatzak erantzuteko edo informazioaren sintesia lortzeko (dokumentu anitzen sintesia batez ere). Arazoa areago korapilatzen duen ezaugarri bat eleaniztasuna da: askotan komenigarria edo ezinbestekoa da dokumentu eleanitzen artean bilatzea edo galderak hizkuntza desberdinetan onartzea. Azalduko dugunez helburu berri hauek lortzeko testuen interpretazio semantikoa ezinbesteko bitartekoa da. Edozein kasutan ez da nahastu behar testuen prozesaketa semantikoa eta web semantikoa. Lehena, hemen zabalduko duguna, hizkuntzaren tratamendu

automatikoarekin lotuta dago, bigarrena, berriz, dokumentuen egituraketa eta etiketatzearekin. Dokumentuen azterketa semantikoetan sakontzeko asmotan Kyoto izeneko proiektu bati ekin diogu EHU eta beste unibertsitate zein enpresaren artean.

Bilaketa-teknologia eta bereziki tresna linguistikoek egiten duten ekarpena azalduko dugu hasieran artikulu honetan. Semantikaren azterketaren inguruan arituko gara gero eta, bukatzeko, Kyoto proiektuaren nondik-norakoak azalduko ditugu.

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]. Datu-base 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 zela-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.

Bilaketen teknologia azken urteetako teknologia arrakastatsuena izan da. Interneteko zabalkundearekin batera bilatzaileak eguneroko tresnak dira jende asko eta askorentzat. *Google*[2] eta neurri txikiagoan *Yahoo*[3] eta *bing*[4] dira bilatzaile osoen eta erabilienak. Hizkuntza nagusiak baino ez dituzte kontuan hartzen eta, ondorioz, euskara bezalako hizkuntza flexiboentzat eta merkatu txikia dutenentzat arazoak daude ohiko bilatzaileetan. 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

[1] http://ixa.si.ehu.es
[2] www.google.com
[3] www.yahoo.com
[4] www.bing.com

(euskarako *elebila*[5]) [4] dira aukerak. 1. irudian bilatzaile horien adibideak ikus daitezke.

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-ren arrakastaren hasierako gakoa). Informazio hori hipertestuen topologiaren azterketaren bidez lortzen da. Hala ere, azken urteetan bilatzaileen arteko lehiakortasuna dela-eta, bilatzaileen barne-funtzionamenduari 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 (horri *search spam* deitzen zaio) eta bilatzaileek publiko ez diren neurriak hartzen dituzte horren aurka.

Emaitzak ebaluatzerakoan bi neurri erabiltzen dira doitasuna (*precisión*) 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 arazotsuak 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.

Dena den informazio-beharra dokumentu edo paragrafo baten eskurapenetik hara doanean gaurko teknologia ez da aski, bilaketaren eta inguruko arloetan azken urteetan aurrerakuntza nabarmenak egon diren arren. Gaiarekin lotuta honako teknika 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

[5] www.elebila.com

ezinbesteko tresna izanik. Emaitzak oraindik ez dira oso ikusgarriak baina ikerketa handia egiten da arlo honetan.

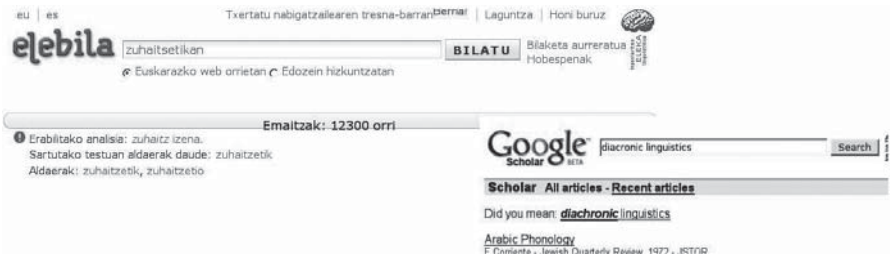
— Bilaketa multimodala: digitalizazioa dela-eta, dokumentuak bilatzeaz gain, argazkiak, irratiko edo telebista/bideoko programak bila daitezke. Bilaketa-sistema hauek metadatuetan (fitxa dokumentala) oinarri daitezke edo bestelako tekniketan (inguruko testua argazkietarako, ahots-testu bihurketa, etab.); eta askotan metodo konbinatuak dira egokienak. Gai hau artikulu honen esparrutik kanpo geratzen da.
— Dokumentuen sintesia: beharra asetzeko emaitza ez da dokumentu bat, dokumentu batzuetan adierazitakoaren sintesia baizik. Pentsa dezakegu produktu batez galdetzen dugunean dokumentu (eta agian hizkuntza) anitzetan produktu horri buruz esaten diren gauza interesgarrienak lortu nahi ditugula.

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.

Morfologian oinarritutako tresnak

Oinarrizko teknologia hitza-lema erlazioa bilatzea da. Analisia esaten zaio hitzetik lema edo forma kanonikoa (hiztegiko sarrera) lortzen duen eragiketari, eta sorkuntza lematik bere forma posible guztiak lortzen dituenari. Flexio handiko hizkuntzetan tratamendu morfologikoa ezinbestekoa da estaldura oso txikia ez izateko.



1. irudia

Galdera zuzentzeko proposamenak

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 linguistikoaz egin daiteke (lemak, aurrizkiak, atzizkiak, paradigmak, aldaketa fonologikoak, etab. erabiliz), edo hurbilpen sinpleago batez, 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.

Aipatutako *elebila* bilatzaileak morfologia integratuta du 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.

Sintaxian oinarritutako aplikazioak

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

Horren adibideak dira www.clusty.com, www.quintura.com, www.kartoo.com webguneak. Horietako batzuetan emaitzak erabiltzen dira bistaratze bereziak egiteko (multzoak, erlazioak...).

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

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 daitezkeen 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[6]. Kalitatea oraindik ez da egokia itzulpen profesionalean erabiltzeko baina halako aplikazioetan erabil daiteke. 2. irudian adibide bat ikus daiteke.

2. irudia

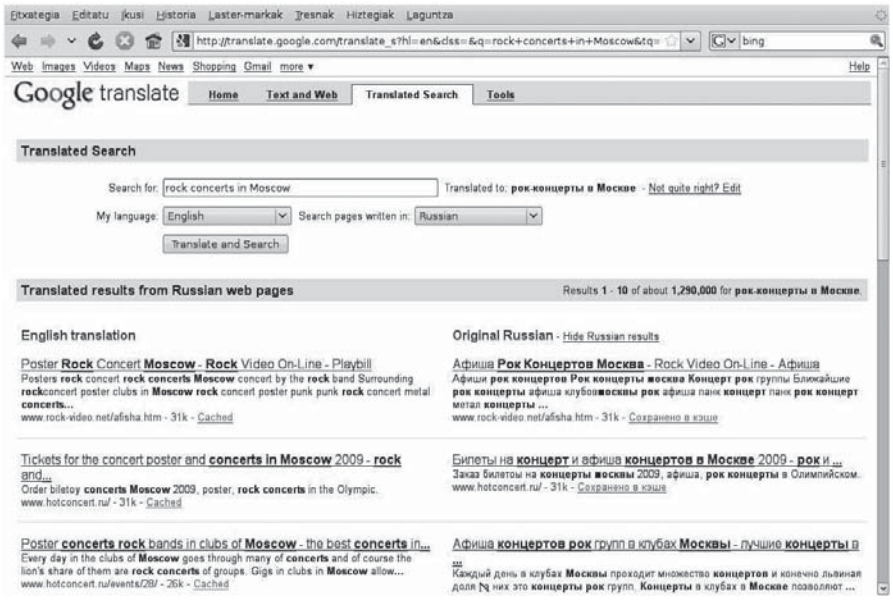
Opentrad itzultzailearen adibidea

CLIR egiteko hainbat diseinu egin daitezke, bilduma hizkuntza bakar batean egotea, baina galderak edozein hizkuntzatan 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 hainbat aukera daude, 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.

---

[6] www.opentrad.com



3. irudia

Translated Search funtzioaren adibidea

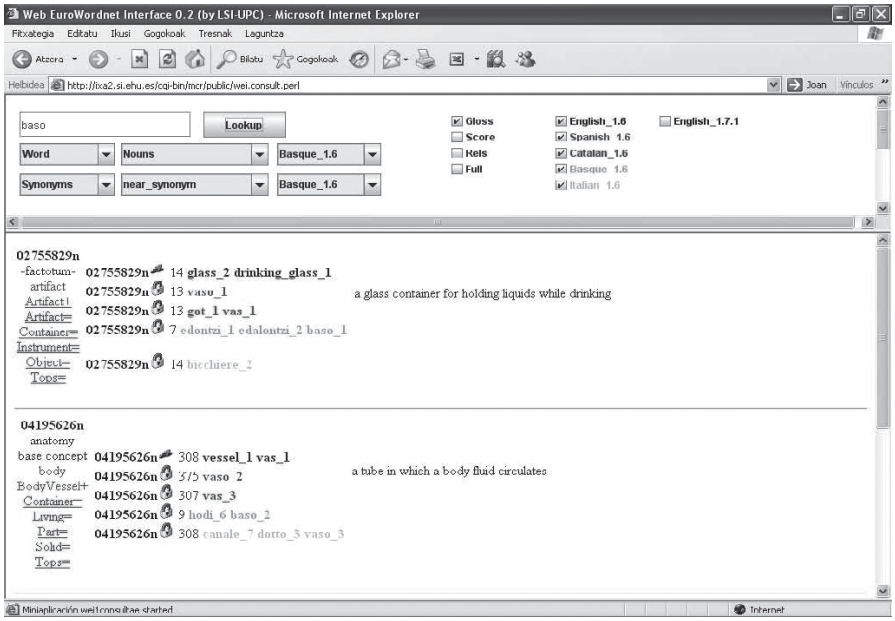
Eleaniztasunean oinarritutako aplikazioen adibide gisa 3. irudiko adibidea dugu Googleren *Translated Search* delakoa itzulpen automatikoan oinarrituta dagoena.

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 (*jatetxe* bilatzen badugu, *sagardotegi* edo *erretegi* hitzak azaltzen diren dokumentuak ez genituzke lortuko). Bilaketa kontzeptuen bitartez egingo balitz, bi arazo horiek arinduko lirateke. Problema bat dago, tratamendu semantikoa nahikoa zehatza ez bada estaldura handitu arren doitasuna galdu egiten dela.

Orokorrean kontzeptuen inbentario diren ontologiak eleaniztunak balira, orduan hizkuntzen arteko bilaketak ere hobeto egingo lirateke, kontzeptu gehienak konpartitzen baitira hizkuntzen artean. Web 2.0 delakoan 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, gizakiak egindako dela edateko basoa, edukitzeko ontzi baten propietateak dituela, instrumentu bezala erabili daitekeela, etab.). Hurrengo irudian azaltzen dira ere



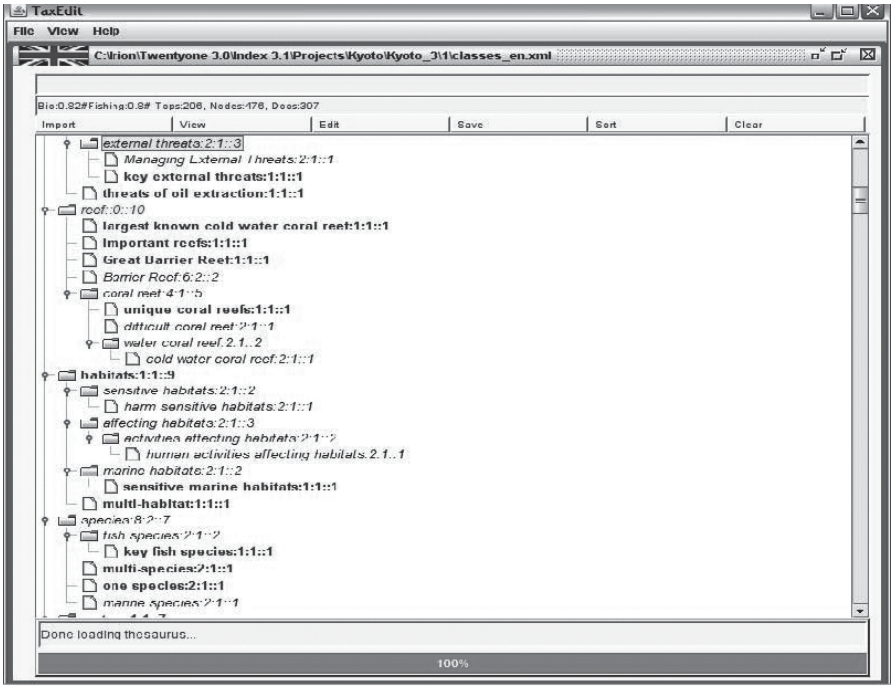
4. irudia

WordNet interfazearen adibidea (*glass-edalontzi, baso*)

KYOTO PROIEKTUA

Asia eta Europako hainbat erakunderen artean garatzen ari den KYOTO proiektuaren (www.kyoto-project.eu) helburua ezagutza modelatzeko eta gertaerak identifikatzeko plataforma bat garatzea da [7]. Garapena komunitatean oinarrituta dago eta hizkuntzen eta kulturen artean erabilgarri izan dadin diseinatu

da. Wiki sisteman oinarritzen den plataformaren bitartez erabiltzaileek domeinu zehatzetako terminologia adostu dezakete, terminoen erauzketa dokumentu eleanitzetatik modu automatikoan lortu eta gero. Adituek termino horiek aldatu, aberastu eta erlazionatu ditzakete. Programen eta adituen lanaren ondorioz ezagutza-egitura konplexuak sortzen dira, hizkuntzarekiko neutralak direnak eta ezkutuan geratzen direnak, baina erabil daitezkeenak testu-bilduma berriak ustiatzeko eta bertatik gertaerak inferitzeko.



5. irudia

Tybot-aren emaitza

Gertaeren datu-base bat sortuko da erabileraren ondorioz eta gertaera horietan zehar bilaketak egin eta nabigatu ahal izango dute erabiltzaileek. Sistema zazpi hizkuntzatarako garatzen ari da eta ingurumena da landutako domeinua, baina beste hizkuntzatarako edo domeinutarako da hedagarria.

Hizkuntzen eta kulturen arteko elkarreragiketa lortzeko partekatutako ontologia bat erabiltzen da, zeinaren bitartez hitzak eta espresioak estekatzen diren. Ontologia errepresentazio formal eta hizkuntzatik independentea da, inferentziarako zein arrazoinamendurako erabil daitekeena. Wiki ingurune baten bitartez lankidetza tresna bat garatu da erabiltzaileek erauzitako informazioa gainbegiratu eta edita de-

zaten. Informazioaren erauzketan aurretik definitutako ontologia eta berari estekatutako espresioak erabiltzen dira patroi ontologiko izeneko predikatuak erabiliz.

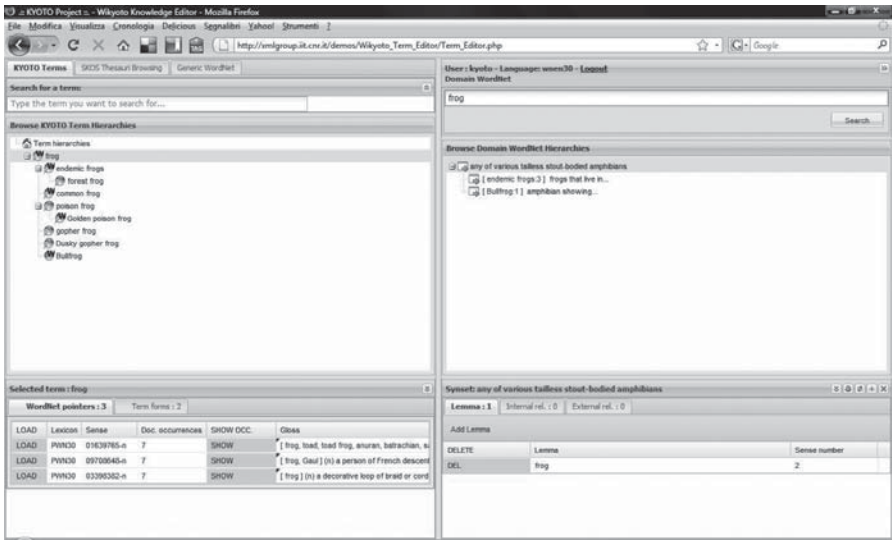
Sistema martxan jarri baino lehen adituen informazio-beharrak aztertu dira. Hauexek dira adibideetako batzuk (ingelesez):

— *Which are the most suitable areas in Europe for pro biodiversity business?*
— *What are the key biodiversity indicators in a certain area?*
— *What is the effect of hedgerows on air quality?*
— *What is the impact of dogs on wildlife?*
— *Are there huge negative effects with regard to eco-networks and alien invasive species?*

Ikus daitekeenez espresioetan termino konplexuak (*pro biodiversity business, biodiversity, eco-networks, alien invasive species*) eta erlazio kausalak (*indicators, impact, effect*) agertzen dira. Ohiko bilatzaileetan arazoak daude galdera hauei erantzun gisa emaitza interesgarriak lortzeko. Kyoto proiektuaren helburua horixe da, gai zehatz baten inguruko galdera konplexuei erantzun egokia ematea. Horretarako hiru helburu lortzeko lanean gaude:

Osatu den testu-bildumatik terminologia eta kontzeptuak hartzen dira automatikoki. Hau egiteko *Tybot* izeneko osagaia eraiki da.

Wikyoto izeneko wiki plataforma bat eratu da adituek hizkuntza eta kulturen artean terminoen eta kontzeptuen esanahia adostu eta argi dezaten. Plataforma honen bitartez adituek terminoen hautapena egin eta sare semantikoan kokatu dezakete, geroago ezagutza hori partekatutako ontologian integra dadin.



6. irudia

Wikyoto-aren itxura

Testu-erauzketarako tresna eleanitz ahaltsu batek terminologia eta ontologia erabiltzen du testu bildumetatik informazio eta gertaera esanguratsuak lortzeko. Aurrekoaren ondorioz datu-bilduma partekatu bakarra sortzen da, bertan bilaketen eginbeharra modu adimentsuan burutu ahal izateko. Tresna honi *Kybot* deritzo.

Bistan da urrats horietan semantika gunea dela. Semantikan oinarritutako prozesaketa aurrera eramateko lau osagai integratzen dira:

— Hizkuntza/kulturari dagokion Wordnet-a, terminoak eta erlazioak jasotzen dituena.
— Termino horien definizioa ontologian, kontzeptu abstraktuak terminoekin lotzeko.
— Kontzeptuen definizioa, hizkuntzatik eta kulturatik neutrala dena.
— Hizkuntzarekin lotutako terminoak eta hizkuntzatik independenteak diren kontzeptuak lotzeko moduaren definizioa.

Gaur egun proiektua erdibidean dago eta oraindik ez dago emaitzarik. Edozein kasutan aipatutako elementu nagusien prototipoak eraiki ditugu eta test fasean daude.

ONDORIOAK

Artikuluan azaldu dugunez, teknologia linguistikoak, tartean semantika, ekarpen handia egin dezake dokumentu digitalen bilaketen arloan. Bilatzaile orokor ezagunenetan oraindik gutxi erabiltzen badira ere, erabilera handitzen doa, orokorrean estaldura handitzeko oso tresna interesgarriak direlako. Tresna linguistiko hauek (sintaxia, itzulpena, semantika) ez dira maizago 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 gainditzeko helburuarekin hainbat ikerketa egiten dira, tartean Kyoto proiektua, zeinaren bitartez ezagutza sakona behar duten informazio-behar konplexuei erantzun egokia eman nahi zaion.

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

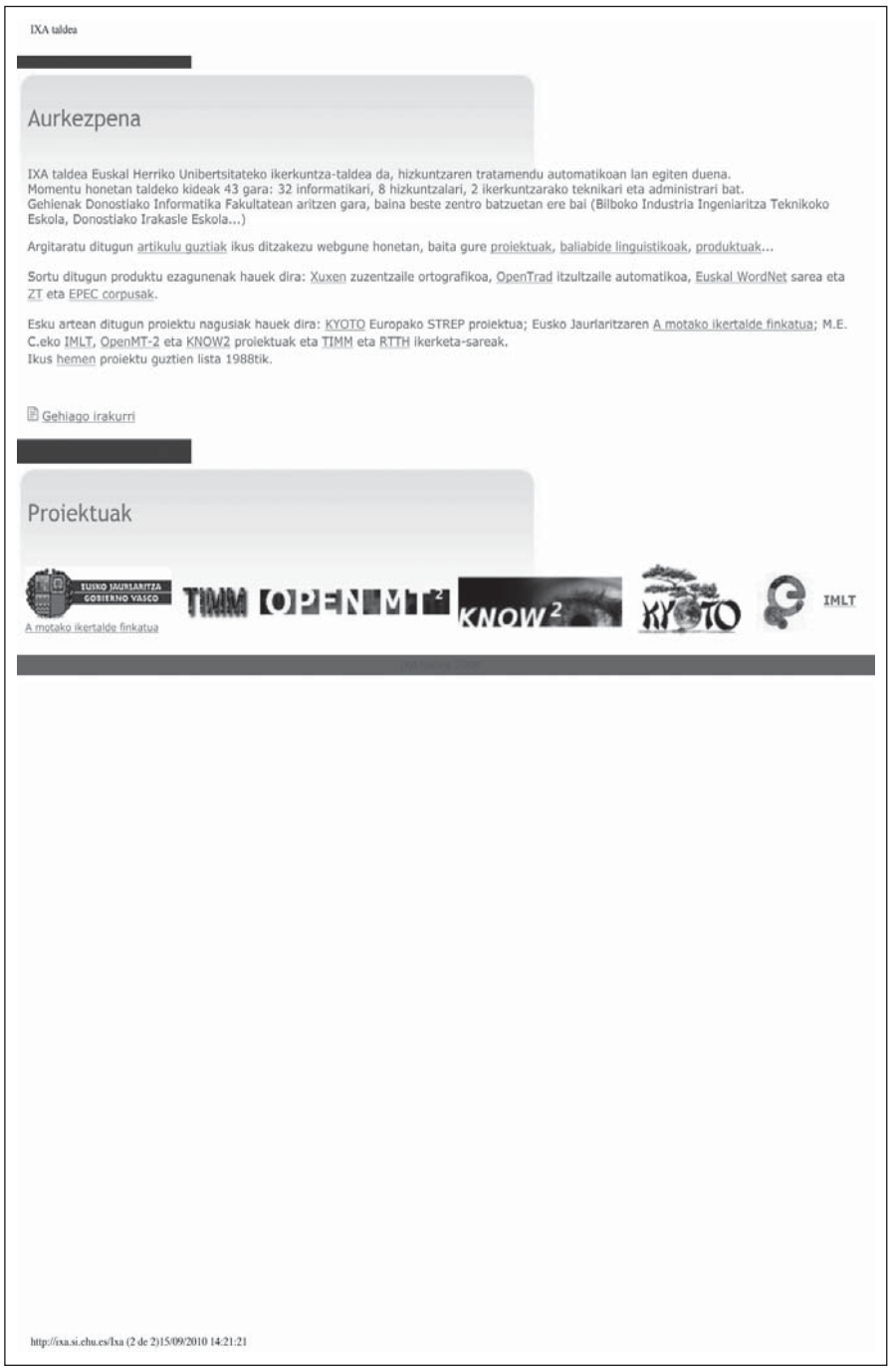
978-84-690-6978-3. http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1190627800/publikoak/pdf

[5] Mayor, A. 2007. *MATXIN: Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz*. Doktoregotesia. 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

[7] Vossen P., E. Agirre, F. Bond, W. Bosma, C. Fellbaum, A. Hicks, S. Hsieh, H. Isahara, Ch. Huang, K. Kanzaki, A. Marchetti, G. Rigau, F. Ronzano, R. Segers, M. Tesconi: «KYOTO: a Wiki for Establishing Semantic Interoperability for Knowledge Sharing across Languages and Cultures», in: *Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models*. IGI Global USA.


IXA taldea
Lang: eu | en
Nor gara?
Etxera
Aurkezpena
Kideak
Zer egiten dugu?
Ikerlerroak
Argitalpenak
Produktuak
Proiektuak
HAP masterra
Beste batzuk
Estekak
Lanpoltsa
Karrera bukaerako proiektuak
Pribatua
HIZKUNTZAREN AZTERKETA ETA PROZESAMENDUA MASTERRA
Kontaktua
Ixa Taldea
649 Posta kutxa
20080 Donostia
ixa@ehu.es
Euskal Herriko Unibertsitatea
ixa
IXA taldea
Hizkuntzaren prozesamendua
Bilatu nahi duzun
Bilatu
Hemen zaude: Ixa
PIL-PILEAN
Udako ikastaroa: Hizkuntzen kudeaketa mundu global batean 2010-Uzt-01
Laponiako samiera hizkuntza eta euskara lankidetzan aztertzen 2010-Eka-28
Hizkuntzaren Azterketa eta Prozesamendua. Master ofiziala 2010-2011 2010-Eka-18
CLARIN proiektuaren bilera Euskal Herriko agenteekin 2010-Eka-14
Berri gehiago ikusi
http://ixa.si.ehu.es/Ixa (1 de 2)15/09/2010 14:21:21


IXA taldea
Aurkezpena
IXA taldea Euskal Herriko Unibertsitateko ikerkuntza-taldea da, hizkuntzaren tratamendu automatikoan lan egiten duena.
Momentu honetan taldeko kideak 43 gara: 32 informatikari, 8 hizkuntzalari, 2 ikerkuntzarako teknikari eta administrari bat.
Gehienak Donostiako Informatika Fakultatean aritzen gara, baina beste zentro batzuetan ere bai (Bilboko Industria Ingeniaritza Teknikoko Eskola, Donostiako Irakasle Eskola...)
Argitaratu ditugun artikulu guztiak ikus ditzakezu webgune honetan, baita gure proiektuak, baliabide linguistikoak, produktuak...
Sortu ditugun produktu ezagunenak hauek dira: Xuxen zuzentzaile ortografikoa, OpenTrad itzultzaile automatikoa, Euskal WordNet sarea eta ZT eta EPEC corpusak.
Esku artean ditugun proiektu nagusiak hauek dira: KYOTO Europako STREP proiektua; Eusko Jaurlaritzaren A motako ikertalde finkatua; M.E. C.eko IMLT, OpenMT-2 eta KNOW2 proiektuak eta TIMM eta RTTH ikerketa-sareak.
Ikus hemen proiektu guztien lista 1988tik.
Gehiago irakurri
Proiektuak
EUSKO JAURLARITZA GOBIERNO VASCO
A motako ikertalde finkatua
TIMM
OPENMT2
KNOW2
KYOTO
IMLT
http://ixa.si.ehu.es/Ixa (2 de 2)15/09/2010 14:21:21