VI

Sarean Zer? Interneteko berriak lotzeko baliabidea

Gregorio Hernández

Eleka Ingeniaritza linguistikoa
Eleka ingeniería lingüística

Laburpena: Interneten bidez hedatzen diren mota desberdinetako informazioek ez daukate elkarren arteko erlaziorik, hau da, erabilera askori begira ez daude ondo antolatuta. Horregatik, informazioa ikuspegi semantiko orokor batetik atzitu nahi dugunean ezinezkoa gertatzen zaigu. Sarean Zer? izeneko produktuak arazo horri aurre egin nahi dio egunkari digitalen esparruan. Bere helburua egunkari desberdinek argitaratutako albisteak lotzea da.

Resumen: Los diferentes tipos de informaciones que se difunden a través de Internet no tienen ninguna relación entre sí, es decir, no están bien organizadas, en lo que a utilización múltiple se refiere. Por eso, cuando queremos acceder a la información desde un punto de vista semántico general, nos resulta imposible. El producto denominado Sarean Zer? Pretende hacer frente a este problema en el ámbito de los periódicos digitales. Su objetivo es relacionar las noticias publicadas por los diferentes diarios.

1. SARRERA

Gaur egun, Interneten argitaratzen den informazioa; blogetako sarrerak, albisteak, eztabaidaguneak eta abarrek ez daukate inolako elkarren arteko erlaziorik, hau da, erabilera askori begira ez dago ondo antolatuta. Horren ondorioz, informazioa ikuspegi semantiko orokor batetik atzitu nahi dugunean ezinezkoa gertatzen zaigu.

Sarean Zer? produktuarekin, problema horri aurre egin nahi diogu, baina esparru mugatu batean. Zehazki, egunkari digitalen esparruan. Esparru zabal eta emankorra non informazioaren nolabaiteko antolaketak prentsa irakurtzeko modu berriak eta aberatsagoak eskaini ditzaken.

Aurkezten den sistemaren helburu zehatza egunkari digital batean argitaratutako albisteak beste egunkari digital batzuetako albisteekin (elkarren antzekotasun mailaren arabera) lotzea izango litzateke. Era horretan, irakurleak albisteak irakurri ahala bere aurrean duen albistearekin erlazionatutako beste albiste batzuk eskuragarri lituzke hizkuntza eta egunkari desberdinetakoak

direla ere. Funtzio honen abantailak ugariak dira: informazioa kontrastatzeko aukera, albiste bateko gai batean sakontzeko posibilitatea...

Sarean Zer? *Le Journal du Pays Basque* kazetan martxan dago eta euskaraz zein frantsesez argitaratzen diren berriak beste hainbat agerkari digitalen antzeko edukiko berriekin lotzen ditu. Hots, irakurleak http://www.lejpb.com webguneko berriak bistaratzen dituenean, berri horren gaiarekin beste agerkari batzuetan argitaratutako berriak lotuta aurkitzen du. Momentuan bost hizkuntza desberdinetako albisteak prozesatzeko gai da: euskara, gaztelania, frantsesa, ingelesa eta katalana.

Arakatzen diren agerkari digitalen zerrendan gertuko egunkari nagusiak (*Gara*, *Diario Vasco*, *Berria*, *El Correo*, *Noticias de Navarra*, …) eta urruneko beste batzuk (*Le Monde*, *The Guardian*, …) daude.

Sistemak administratzaileak ezarritako zerrenda bateko egunkari digitaletara periodikoki konektatzen da eta *Le Journal*eko albisteen antzekoak identifikatzen ditu. Elkarren antzekoak sistemaren datu-basean gordeko dira *Le Journal*eko web-gunetik kontsultagai egon daitezen momentu oro.

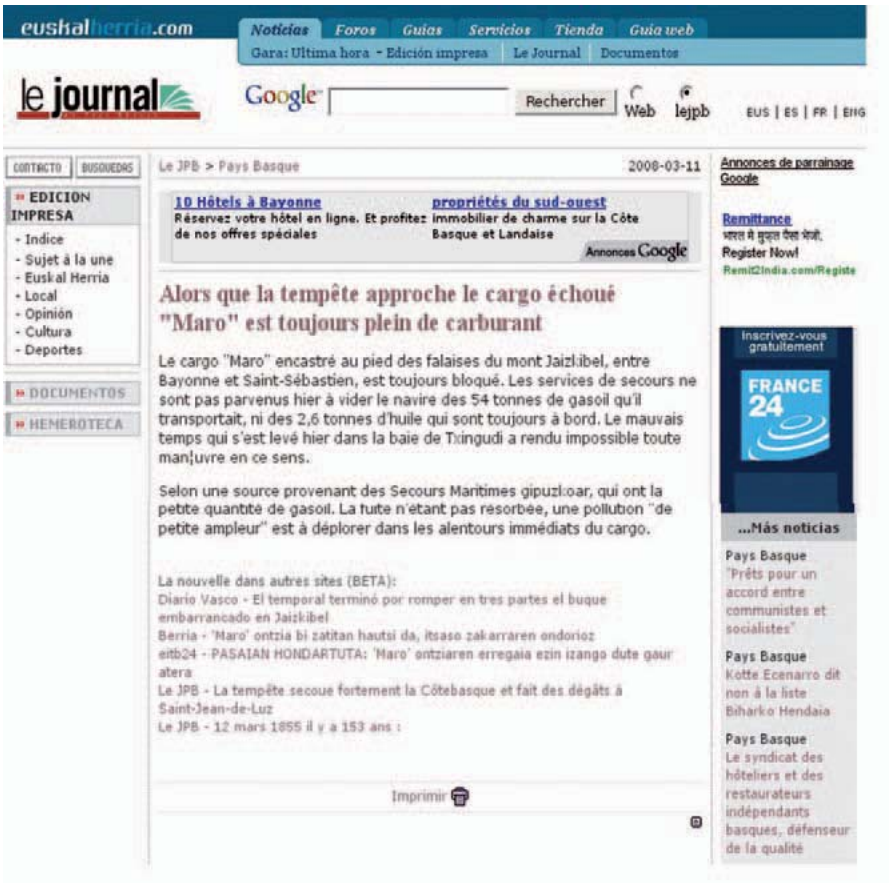
2. TEKNIKAK (ARTEAREN EGOERA)

Informazio eleaniztunaren tratamenduari lotutako ikerketa-gaien artean, *Cross-Language Information Retrieva*l (CLIR) [1] [2] eta horren baitan dagoen *Cross-language Similarity* ditugu. Eleaniztasunaren eremura eramandako teknologia asko bezala, horiek ere elebakarrean egindako lanetik abiatzen dira. Bai elebakarrean, bai eleaniztasunaren eremuan egindako lanak ugariak dira.

Dokumentuen arteko erlazioak bilatzen duen antzekotasun semantikoa neurtzeko metodoak bi ezagutza motatan oinarritzen dira: ezagutza linguistikoan eta metodo estatistikoetan [3]. Beraz, ezagutza horiek nola aplikatzen diren, metodoen arteko ezberdintasunak sortzen dira. Jarraian, gai honen arazo nagusiak eta literaturan aurkitu ditugun metodoak azalduko dira.

Dokumentuen arteko antzekotasunaz aritzen garenean, ez dugu zehazten zeren arabera egiten dugun konparazioa. Objektuen arteko antzekotasuna definitzerakoan, hainbat ezaugarri har daitezke kontuan:

— Ezaugarri estilistikoak; esaldien luzera, hitzen aldaerak…
— Metainformazioarekin erlazionatutako ezaugarriak: sorrera-data, egilea, formatua…
— Ezaugarri semantikoak, hau da, hitzek edo lemek eskaintzen duten informazio bera, gehi haien arteko erlazio semantikoak, askotan ontologia baten bidez adierazia. Ataza horretarako elementu batzuk giltzarri izan daitezke: entitateak, terminologia…



1. irudia

Le Journal-eko berri bat eta Sarean Zer? sistemak
beste hedabideetan topatu dituen erlazionatutako berriak

Informazio horiek guztiak konbinatu daitezke, eta, adibidez, izenburuen azterketa sintaktikosemantikoaren bidez ezar liteke dokumentuen arteko antzekotasunerako oinarrizko analisia.

Antzekotasuna neurtzeko prozesuaren lehenengo pausuan, dokumentua eredu konputagarri bihurtu behar da. Ezinbesteko urratsa da programazio-lengoaien bidez prozesatzeko Modelizazio horretan bi ezaugarri zaintzen saiatzen da: alde batetik, adierazgarritasuna, hau da, adierazi nahi dugun kontzeptuaren ezaugarri zuzenak kontuan hartzea eta zarata baztertzea; eta, bestetik, eraginkortasuna, hau da, konputatze-prozesuaren abiadura ahalik eta azkarrena

izatea. Beraz, ezaugarrien hautaketa eta nola egituratzen den oso urrats garrantzitsua da modelizazioa egitean.

Antzekotasun semantikoaren arloan, hainbat eredu planteatu dira. Batzuek ezagutza linguistikoari garrantzi handiagoa ematen diote, eta besteek, aldiz, estatistikoari. Azken multzo horretan, ezagutza estatistiko hutsa ibiltzen duten ereduak daude. Dokumentuen edukiak adierazteko, bektoreak erabiltzen dira, eta horietan egituratzen dira testuko hitzak, gako-hitzak (*keyword*) ego n-gramak eta horien maiztasunak bektore batean egituratuta adierazten dira dokumentuen edukiak.

Hala ere, bada adierazgarritasun estatistikoari eragiten dion fenomeno bat hizkuntz guztietan gertatzen dena baina batez ere eranskarietan, adibidez, euskaraz, nabarmentzen dena: flexioa. Fenomeno honen ondorioz, lema berak hainbat testu-forma har ditzake (sare, sarea, sareengatik, sareekin…) eta agerraldi horiek lemarekin erlazionatu ezean, barreiadura sortzen da datu estatistikoetan. Beraz, ezagutza linguistikoa aplikatzea (kasu honetan, lematizazio prozesu bat) behar-beharrezkoa da. Beste eredu batzuetan, terminoen bidez adierazten da dokumentuen edukia. Horrelakoetan, termino-erauzleak erabiltzen dira, eta erauzketa prozesuan eredu morfosintaktikoak aplikatzen dira. Horretarako. lematizazioaz gain, kategoriaren eta azaleko gramatika-egituraren informazioa ere behar da.

Orain arte aipatu ditugun ereduak informazio lexikoaz baliatzen dira. Beste eredu batzuk, ordea, horrelako informazioaz gain, esaldiaren egitura kontuan hartzen dute [4]. Horretara, hitzen testuinguruak alderatu daitezke. Esaldiak eta lexikoa (terminologia) grafo orientatuen bidez adierazten dira.

Dokumentu eleaniztunen kasuan, kontzeptuak hainbat hizkuntzatan adierazita daude, eta, lehen azaldutako ereduak hizkuntzatik independenteak ez direnez, ezin dira zuzenean erabili. Hori dela eta, lan askotan planteatzen den irtenbidea testuak itzultzea da. Testuak hizkuntza berean daudelarik, konparagarriak dira informazio elebakarrerako planteatu ereduetan. Hala ere, itzulpen-eredua erabiltzeak erroreak eragiten ditu, gehien bat lortutako itzulpenak ez direlako guztiz zuzenak. Itzulpen-eredua garatzeko ere hainbat teknika ikertu dira. Oinarrizkoena hiztegi elektronikoetan oinarritutako itzulpen-eredua litzateke [3]. Eredu horretan, hiztegian dauden sarrerak itzultzen dira desanbiguatze-prozesurik aplikatu gabe. Desanbiguaziorik eza leuntzeko, hiztegi probabilistikoetan oinarritutako ereduak erabil daitezke [5]. Eredu honetan desanbiguazio orokorra izan arren, itzulpenak hobetu egiten dira. Dena den, desanbiguazio-maila handia lortzeko, itzulpen automatikorako garatzen diren sistemak erabili beharko lirateke.

Lehen esan bezala, dokumentu elebakarrak modelatzeko erabiltzen diren ereduak aplikatzeko, itzulpen-eredu baten bidezko aurreprozesua egin behar da. Urrats hori ez da beharrezkoa hizkuntzatik independenteak diren ereduak erabiliz gero. Eredu hauek gai dira hizkuntza jakin batzuetan emandako dokumentuen edukiak era normalizatu batean adierazteko. Beraz, ez dute itzulpen-eredurik behar. Hala ere, mota honetako ereduetan modelizazio-prozesua

konplikatu egiten dira. Thesaurus eleaniztunetan oinarritzen diren ereduetan adibidez, bi urratsetan egiten da modelizazio-prozesua [6]. Lehendabizi, eskuz etiketatutako testu batzuk trebatze-lagintzat erabiliz, thesaurus-eko kontzeptuen eta lemen arteko erlazioak aztertu eta gordetzen dira. Erlazio hori estimatzeko, egiantz-arrazoia (log-likehood ratio) eta kideko elkartze-neurriak erabiltzen dira. Ondoren, testu berriei thesauruseko kontzeptuak automatikoki esleitzen zaizkie. Erlazioak eta pisuak bektore batean gordetzen dira. Beste metodo bat da thesaurus espezifiko bat erabili ordez WordNet-ez baliatzea [7]. Kasu horretan, kate lexikalak (lexical chains) erabiliz adierazten da dokumentuaren edukia [8]. Kate lexikal batean erlazio semantikoa dituzten hitzak (terminoak) gordetzen dira.

Orain arte dokumentuak eredu konputagarrien bidez nola adierazten diren aritu gara. Hurrengo pausua dokumentu modelizatuen arteko antzekotasuna neurtzeko teknikak aztertzea izango litzateke.

IR (Informazioaren berreskurapena) arloko teknikak erabili ohi dira antzeko dokumentuak bilatzeko. Bektore-espazio ereduan (vector space) oinarritutako neurriak dira ohikoenak, bektoreen arteko kosinua kalkulatzen dutenak hiru neurri desberdinak erabiliz bektoreen osagaiak adierazteko: bitarra, maiztasuna, eta LSA (Latent Semantic Analysis) transformazioa. Zabalduena maiztasuna erabiltzea da, baina LSA bidez informazio semantiko inplizitua atzeman daiteke eta oso emaitza onak lor daitezke [9].

3. SISTEMAREN ARKITEKTURA

Sistemaren arkitektura hiru modulo nagusik osatzen dute: albisteen web-arakatzailea, albisteen edukien modelatzailea eta elkarren antzekoen detektagailua. Labur azalduta, web-arakatzaileak egunkari digital jakin batzuetan argitaratzen diren albiste berriak jaso eta testu-formatura pasatzen ditu. Ondoren, modelatzaileak teknika linguistiko eta estatistikoez baliatuta testuak Bektore Espazioaren ereduaren bidez adierazten ditu. Azkenik, behin arakatzaileak jasotako albiste guztiak Bektore Espazioan adierazi eta gero elkarren antzekoen bilaketari ekingo zaio.

— **Albisteen web-arakatzailea**: Albisteak hainbat prentsa web-gunetatik hartzen dira. Zentzu horretan, web arakatzaileak web-guneak bisitatu eta RSS jarioen bidez argitaratutako dokumentu berrien izenburuak, sarrerak eta HTML orria jasotzen ditu. HTML dokumentutik albistearen edukia bakarrik erauzteko helburuarekin *HTML-wrapperrak* garatu dira. Era horretan, edukiaren parte ez diren testuak (Menuak, iragarkiak...) baztertzen dira. Wrapperren diseinuan Perl programazio lengoaia erabili da. Momentuan, arakatzen diren hedabideen zerrenda honakoa da:

  - *El Mundo* - www.elmundo.es
  - *ABC* - www.abc.es

- *El Periódico de Catalunya* - www.elperiodico.com
- *La Vanguardia* - www.lavanguardia.es
- *El País* - www.elpais.es
- *Vilaweb* - www.vilaweb.cat
- *Diario Vasco* - www.diariovasco.com
- *El Correo* - www.elcorreodigital.com
- *Gara* - www.gara.net
- *Berria* - www.berria.info
- Noticias taldea - http://www.noticiasdenavarra.com
- eitb24.com
- *Le Monde* - www.lemonde.fr
- *Liberation* - www.liberation.fr
- *NY Times* (USA)- www.nytimes.com
- *The Times* (UK) - www.timesonline.co.uk
- *BBC News* (UK) - news.bbc.co.uk
- *Guardian* (UK) - http://www.guardian.co.uk/
- *Al Jazeera* (Arabe) - english.aljazeera.net
- *La jornada* - www.jornada.unam.mx

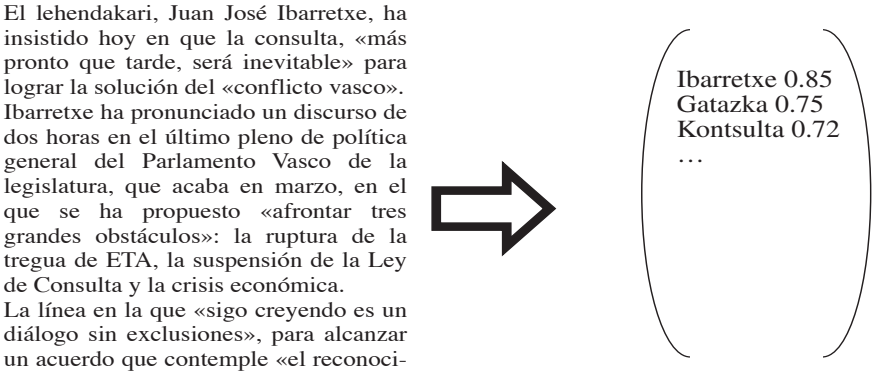
El lehendakari, Juan José Ibarretxe, ha insistido hoy en que la consulta, «más pronto que tarde, será inevitable» para lograr la solución del «conflicto vasco». Ibarretxe ha pronunciado un discurso de dos horas en el último pleno de política general del Parlamento Vasco de la legislatura, que acaba en marzo, en el que se ha propuesto «afrontar tres grandes obstáculos»: la ruptura de la tregua de ETA, la suspensión de la Ley de Consulta y la crisis económica.
La línea en la que «sigo creyendo es un diálogo sin exclusiones», para alcanzar un acuerdo que contemple «el reconocimiento de la existencia del pueblo vasco y el derecho a decidir su propio futuro«
…

2. irudia

Arakatzaileak berriak jasotzen ditu hainbat hedabidetatik eta wrapper-ak erabilita, berriaren testua bakarrik berreskuratzen du

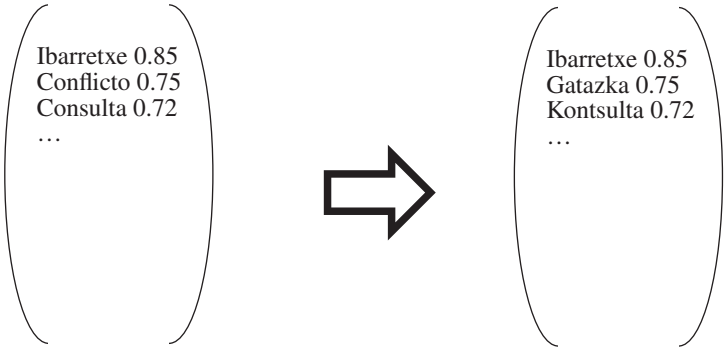
— **Edukien modelatzailea**: Hainbat teknika edo paradigma daude testuen modelizazioaren alorrean. Guk eskura geneuzkan baliabideak eta era-

ginkortasuna kontuan hartuta Bektore Espazio eredua erabili dugu. Bertan testu-edukia gako hitzak eta beren adierazgarritasun-mailak dituen bektore baten bidez adierazten da. Hitz-gakoak edo lexiko adierazgarria identifikatzeko teknika linguistikoez baliatu gara. Zehatz mehats, izenak, aditzak, adjektiboak, entitateak eta hitz anitzeko terminoak hartu dira hitz-gako posibletzat. Izenak, aditzak eta adjektiboak identifikatzeko **Eustagger** POS/lematizatzailea erabili da euskararen kasuan. Ingelesa, katalana eta gaztelaniaren kasuan **Freeling** etiketatzailea erabili da. Azkenik, frantseserako **Apertium**-eko etiketatzailea erabili da. Entitateak eta hitz anitzeko terminoak harrapatzeko ordea aplikazio berriak garatu dira. Entitateen harrapatzailea ahalik eta zehatzena izateari, hau da, doitasunari eman diogu lehentasuna. Batez ere, okerreko entitateek eredua distortsiona dezaketelako. Gainera, testu periodistikoetan entitateek oso unitate adierazgarriak dira. Beraz, haien detekzio zehatzak eragin zuzena dauka antzeko dokumentuak identifikatzerakoan. Terminoen harrapatzailearen kasuan ere doitasuna izan da gure lehentasuna. Zentzu horretan, detekzio-prozesua hiztegi terminologikoetan oinarritu da. Bestalde, hizkuntza desberdinetako dokumentuen bektoreak elkarrekin alderatu ahal izateko hiztegien bidezko estrategia jarraitu da. Dena dela, bektore bat beste hizkuntza batera pasatzean haren errepresentazioa kaltetu egiten da. Alde batetik, gako-hitz batzuk ez direlako itzultzen (ez daude hiztegian), eta bestetik itzulpenetan anbiguotasunak gertatzen direlako. Hiztegien estaldura hobetzeko ere kognaduran oinarrituko teknikak aztertu dira. Modelizazioaren bukaeran hitz-gakoen pisua kalkulatzeko tfi-idf multzoko neurrien bidez egin da.



3. irudia

Testua modelizatu egiten da eta bektore bat lortzen da hitz-gakoekin eta hauen garrantziarekin



4. irudia

Bektoreak euskara itzultzen dira, hurrengo pausuan hizkuntza ezberdinetako berrien arteko antzekotasuna kalkulatu ahal izateko

— **Elkarren antzekoen detektagailua**: Bi dokumenturen arteko distantzia estimatzeko dokumentuen bektoreen arteko kosinua kalkulatzen da. Hau da, bi bektoreak normalizatu eta geroko biderketa eskalarra. Gutxieneko balio bat gainditzen dutenak elkarren antzekotzat hartuko dira. *Le Journal*-eko berri batek antzeko dokumentuak baldin baditu, hauen estekak aterako dira, 1. irudian agertzen zen bezala.

[1] Modern Information Retrieval. Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Addison-Wesley. 1999.
[2] CLIR: Cross-language information retrieval. G Grefenstette-Kluver, Boston, 1998.
[3] B.MATHIEU, R.BESANÇON, C.FLUHR, «Multilingual document clusters discovery», *RIAO* 2004, Avignon, France, April 26-28, 2004.
[4] KHALED M. HAMMOUDA «Efficient Phrase-based document indexing for web document clustering».
[5] TIM LEEK, HUBERT JIN, SREENIVASA, RICHARD SCHWARTZ «The BBN crosslingual topic detection and tracking system», BBN Technologies, MA.
[6] RALF STEINBERGER, BRUNO POULIQUEN, JOHAN HAGMAN, «Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC» (2002).
[7] EUROWORDNET: a multilingual database with lexical semantic networks. P Vossen-Norwell, MA: Kluwer Academic Publishers, 1998.
[8] NICOLA STOKES, JOE CARTHY, «Combining Semantic and Syntactic Document Classifiers to improve First Story Detection», Department of Computer Science, University College Dublin, Ireland.
[9] S. T. DUMAIS, T. A. LETSCHE, M. L. LITTMAN, AND T. K. LANDAUER. Automatic cross-language retrieval using latent semantic indexing. Working notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, 1997.