ANUARI DE FILOLOGIA. ESTUDIS DE LINGÜÍSTICA (Anu.Filol.Est.Lingüíst.)
3/2013, pp. 187-201, ISSN: 2014-1408

EL PAPER DE LA LINGÜÍSTICA COMPUTACIONAL EN LA CERCA D'INFORMACIÓ

JORDI ATSERIAS
Fundació Barcelona Media-Yahoo! Research Barcelona
jordi@yahoo-inc.com
HUGO ZARAGOZA[1]
Websays
hugo.zaragoza@websays.com

RESUM

Aquest article presenta una visió general de les noves aplicacions de cerca d'informació que estan emergint fent ús de la semàntica superficial. Es presenten tres exemples d'aquest tipus d'aplicacions desenvolupades per la fundació Barcelona media-Yahoo! Research Barcelona, Yahoo! correlator, TimeExplorer i Quest.

PARAULES CLAU: Recuperació d'informació; cerca; lingüística computacional; semàntica superficial; web semàntica.

THE ROLE OF COMPUTATIONAL LINGUISTICS IN THE SEARCH FOR INFORMATION

ABSTRACT

This paper presents a general view of the new emerging search applications which make use of shallow semantics. Three different applications developed by fundació Barcelona media-Yahoo! Research Barcelona, Yahoo! correlator, TimeExplorer and Quest, are presented.

KEY WORDS: Information Retrival; Search; Computacional Linguistics; shallow semàntics; semantic web.

1. INTRODUCCIÓ

La gent s'ha acostumat a buscar documents al web i ha anat deixat de banda la idea inicial de buscar "informació". El model de cerca per paraules sembla haver arribat al seu límit, i mentre el volum de la informació disponible en format electrònic creix cada cop a més velocitat, la nostra capacitat de cercar i integrar la informació segueix sent molt limitada. El principal entrebanc és que molta d'aquesta informació està poc o gens estructurada (sobretot al web) i l'única manera d'accedir-hi actualment es veu limitada al paradigma de cerca per mots. D'altra banda, aquesta falta d'estructura en las dades fa que la majoria dels models de cerca es limitin a retornar una sèrie de documents.

Com a alternativa a aquesta cerca basada a buscar documents mitjançant paraules, es parla de construir sistemes intel·ligents, els quals "comprendran el llenguatge escrit" i permetran fer "inferències" (amb la complexitat i el cost en

[1] El treball presentat en aquest article es va dur a terme mentre l'autor treballava a Yahoo! Research.

Data de recepció: 02/IX/2013
Data d'acceptació: 05/X/2013

temps i recursos que això representa). En la visió de la cerca semàntica, l'eina de cerca intentarà "entendre" el que se li està preguntant i/o el contingut dels texts.

Encara hi ha molt de terreny a explorar entre aquestes dues visions extremes: les tecnologies de cerca poden ser emprades per explotar informació semàntica superficial i proporcionar noves i poderoses formes de cercar informació. Part d'aquest poder està a trobar-dissenyar-proporcionar les interfícies gràfiques correctes que permetin a l' usuari, d'una manera natural i senzilla, tant especificar que és el que està buscant, com navegar pels resultats, evitant caure en la complexitat d'entendre completament la pregunta o els documents però utilitzant una semàntica "superficial" (a priori molt més robusta, ràpida, escalable) per proporcionar una cerca molt més potent que la cerca basada en paraules. De fet, no és que el programa comprengui realment el que estem cercant, sinó que el programa té la capacitat, mitjançant una sèrie d'algoritmes, d'analitzar el significat de les paraules (o expressions) en el seu context.

Un dels reptes actuals a la web és trobar aquestes noves formes de cercar informació (no documents) i de com la informació semàntica pot ajudar en aquestes tasques. En general l'usuari no està acostumat a poder fer preguntes complexes, ni a formular les seves preguntes en termes lingüístics. En aquest article veurem diferents exemples de com la informació proporcionada automàticament per la lingüística computacional es pot incorporar als motors de cerca, i com es poden desenvolupar noves interfícies de cerca molt més potents que les actuals, que facin un ús implícit d'aquesta informació semàntica sense canviar la manera en què l'usuari està acostumat a fer les cerques.

A diferència d'altres iniciatives de millora de les aplicacions de cerca, les aplicacions com Correlator[2] o Yahoo! Quest[3] no es basen a millorar la comprensió de la pregunta o la comprensió total del documents, sinó en l'ús d'informació semàntica superficial (com ara les entitats) per aconseguir estructurar millor i arribar més fàcilment a la informació.

En aquest marc, la lingüística computacional ens permetrà d'explorar i mostrar noves formes de cerca d'informació sobre col·leccions de gran volum que vagin més enllà del paradigma de cerca de paraules i documents. Quan emprem un cercador al web hem d'expressar els nostres termes de recerca amb paraules clau que apareixeran en els texts que l'eina de recerca ens tornarà com a resultat. Si volem saber quins dinosaures eren propis o s'han trobat a Argentina, haurem d'introduir un sèrie de paraules al cercador com per exemple: Argentina dinosaures. Aquesta cerca pot no trobar els resultats desitjats si en el text no apareixen exactament aquestes paraules (per exemple en lloc de dinosaures apareix velociraptor). Però només amb una mica de coneixement semàntic, que permeti saber que el velociraptor és un tipus de

[2] http://correlator.sandbox.yahoo.com

[3] http://quest.sandbox.yahoo.com

dinosaure o que Argentina és un país, la nostra eina de cerca pot trobar i organitzar molt millor la informació.

L'ús de la lingüística computacional és una alternativa-complement a la visió de la web i de la cerca d'informació de l'anomenada web semàntica. La web semàntica té com a objectiu crear un medi universal per a l'intercanvi d'informació significativa (semàntica), d'una forma comprensible per a les màquines, del contingut dels documents de la web. Però tot i el creixement d'aquest tipus d' informació estructurada al web, la majoria de la informació al web continua sent text no estructurat.

La web semàntica i la lingüística computacional són visions complementàries que es poden ajudar mútuament (per exemple, Mika *et al*. 2008). Els diferents etiquetadors semàntics o la creixent tecnologia capaç d´enllaçar les entitats en un text amb els recursos de la web semàntica ens permetrà cada cop més salvar la distància entre aquestes dues visions.

Les aplicacions de la lingüística computacional a la cerca en grans volums d'informació a la web, encara té però molts reptes. No només poder analitzar texts en diferents llengües sinó també els usos particulars del llenguatge que es donen a la web. L'evolució de la web (Web 2.0) ens ha dut a facilitar la possibilitat que qualsevol usuari publiqui informació (blogs, piulades, etc). Les restriccions d'alguns d'aquest mitjans (per exemple, la restricció de 140 caràcters per piulada a Twitter), ha comportat també un canvi en la manera de fer servir la llengua (per exemple, noves abreviacions, transcripcions fonètiques, variacions no estàndard de l'ortografia). Aquestes noves variacions de la llengua representen un nou repte per a la lingüística computacional (vegeu LREC UGC workshop).[4]

Tot i no aspirar a poder processar tot el text de la web, hi ha certes restriccions necessàries per fer viable el processament de grans col·leccions de documents, per exemple algoritmes de cost lineal, la majoria d'algoritmes d'anàlisi sintàctica són quadràtics en els millors dels casos. L'altre gran repte per a la lingüística computacional és l'escalabilitat. Noves aproximacions, sobretot l'anomenat "Cloud computing" (com ara Map-Reduce a Hadoop)[5] han permès poder escalar de manera fiable el processament massiu de dades.[6]

2. La Semàntica "Superficial"

La lingüística computacional encara està lluny de poder comprendre automàticament el significat d'un document. Però sí que disposem de tecnologies capaces de poder extreure automàticament certa informació semàntica d'un document amb una qualitat i unes restriccions de

[4] http://nlp4ugc.barcelonamedia.org/Site/Welcome.html

[5] http://hadoop.apache.org/

[6] Tot i que la majoria de plataformes de NLP, com UIMA amb UIMA-AS, han seguit diferents aproximacions d'escalabilitat.

temps/recursos acceptables per a poder processar grans volums de texts. Tot i que aquests tipus d'anotacions semàntiques (com per exemple les entitats nombrades) han despertat un gran interès, no hi ha cap consens de com utilitzar aquestes anotacions per millorar els algoritmes i les aplicacions de cerca d'informació.

Els següents apartats descriuran alguns d'aquest elements semàntics bàsics que es poden extreure automàticament i ràpidament d'un text i que són la base de les aplicacions com Correlator o TimeExplorer: les entitats nombrades, les expressions temporals i espacials o la polaritat i l'anàlisi del sentiment.

2.1. Entitats Nombrades

El reconeixement i la classificació d'entitats nombrades, és a dir, noms de persones, llocs, organitzacions, etc., és una tasca en què s'ha vingut treballant durant els darrers anys (per exemple CONLL[7] 2002 i 2003 shared tasks i Evalita[8] 2009-2011) i per la qual s'obtenen resultats més que acceptables en diverses llengües (Nadeau i Sekine 2007).

La possibilitat de detectar aquest tipus d'entitats en un document ens permet abordar altres maneres de proporcionar informació a l'usuari. Per exemple, donada una pregunta, retornar una llista ordenada (rànquing) de les entitats nombrades més rellevants per a la nostra cerca (entity ranking), en lloc de retornar documents (Zaragoza *et al.* 2007). Diverses competicions focalitzen diferents aspectes de la rellevància de les entitats donada una pregunta (Enterprise track TREC 2008 per trobar experts en un tema, INEX Entity Ranking track o l'Entity track del TREC 2009 per trobar les entitats rellevants per a una pregunta).

A més de poder trobar la menció d'una entitat nombrada en un text podem intentar enllaçar-la amb la informació estructurada disponible referent a aquesta entitat. Una mateixa persona o lloc poden compartir nom, acrònim o àlies. Així, per tal de poder enllaçar una menció amb un repositori d'entitats, haurem de decidir a quina entitat concreta es fa menció en el text. Aquesta tasca és coneguda com a Entity Linking, o Entity Dissambiguation. Per poder fer aquesta "desambiguació", de la mateixa manera que per a les paraules, necessitarem un diccionari (repositori) dels possibles sentits (entitats). Per exemple, la Viquipèdia o d'altres recursos de la web semàntica (dbpedia, imdb). Aquesta tasca pot arribar a ser força complexa, bé per l'existència de múltiples entitats amb els mateixos noms o mencions (per exemple, existeixen fins a nou persones[9] amb el nom de michael jordan a la viquipèdia anglesa) o per haver de detectar quan la menció fa referència a una entitat que no existeix en el nostre

[7] http://ifarm.nl/signll/conll

[8] http://www.evalita.it

[9] http://en.wikipedia.org/wiki/Michael_Jordan_%28disambiguation%29

repositori. Normalment el problema a més esdevé més general ja que la majoria d'aquests repositoris no només contenen entitats nombrades sinó que podem tenir entitats com ara "física nuclear". Hi ha diverses competicions (com ara Entity-Linking task TAC 2009 i 2012) que han permès aprofundir en aquest problema i un nombre cada cop més gran d'aplicacions online que permeten aplicar aquests mètodes a un text (TagMe,[10] Zemanta,[11] Spotlight,[12] etc).

A més de trobar les entitats més rellevants i de poder relacionar aquestes entitats amb les nostres bases de coneixement, també voldrem poder explicar per què aquestes entitats són rellevants o trobar un document on una d'aquestes entitats sigui central. En un document hi poden apareixen moltes entitats; de cara a aplicacions de cerca d'informació, també és importat poder determinar quina és l'entitat/s principal/s d'un document (Moens *et al.* 2006).

2.2. Expressions Temporals i Espacials

El temps forma part de manera natural de moltes aplicacions de cerca (Kanhabua *et al.* 2012 i Demartini *et al.* 2010). Molts resultats de cerca tenen un component temporal però són pocs els treballs en recuperació d'informació que fan servir aquesta dimensió més enllà de mostrar els documents més recents. Les aplicacions-necessitats d'informació on el temps és fonamental són moltes. Per exemple, poder contestar com evoluciona un tema al llarg del temps, o com una entitat guanya o perd rellevància en un tema en diferents moments del temps o fins i tot cercar quines previsions estem fent sobre el que succeirà en el futur (Baeza-Yates 2005).

Un document textual pot fer referències més o menys complexes a d'altres moments en el temps. Aquestes expressions temporals són un altre tipus de semàntica superficial que es pot arribar a extreure automàticament, tot i la vaguetat o contextualitat (moltes referències temporals, tals com "avui", poden dependre de la data de publicació del document o d'altres com "algun dilluns" poden ser difícils de manegar per un cercador).

L'estàndard TimeML (Pustejovsky *et al.* 2005) defineix una formalització de la informació que es pot extreure de les expressions temporals. TimeML aborda quatre tipus d'expressions, DATE (unitats iguals o superiors a un dia), TIME (unitats inferior a un dia), DURATION (quan s'indica la duració) y SET (quan s'indica la freqüència d'un esdeveniment). L'existència de corpus anotats amb expressions temporals,[13] TimeBanks, per diverses llengües fa que s'hagi pogut millorar i comparar diferents sistemes existents per a capturar automàticament les expressions temporals, des del precursor TARSQI[14] per

[10] http://tagme.di.unipi.it
[11] http://www.zemanta.com/demo
[12] http://spotlight.dbpedia.org/demo
[13] http://www.timeml.org/site/timebank/timebank.html
[14] http://www.timeml.org/site/tarsqi/

l'anglès, o HeidelTime per l'anglès-alemany-castellà (Strötgen *et al.* 2012), al (Llorens *et al.* 2009) pel català.

Tot i la complexitat d'algunes de les expressions temporals que fem servir en el llenguatge, una gran majoria d'aquestes expressions (especialment les de tipus DATE i DURATION) poden ser resoltes automàticament a una data concreta o a un rang de dates i incorporades a sistemes de cerca, com els sistemes d'informació geogràfica (Guerrero i Saurí 2013) o d'altres, com Correlator i Time Explorer que es descriuen a la secció 3, on es fan servir línies de temps (timelines com simile)[15] per visualitzar millor la informació amb un fort component temporal.

Un altre component de semàntica bàsica per entendre millor el contingut dels documents és el component espacial. La informació continguda a la Viquipèdia o serveis de geolocalització com placeFinder[16] permet associar moltes de les entitats nombrades (com ara Barcelona) a les seves coordenades geogràfiques, i d'altres serveis més sofisticats fins i tot informació (per exemple polígons) que permet delimitar aquestes entitats geogràficament en un mapa. Ser capaç de geolocalitzar entitats pot permetre no només construir aplicacions on el resultat es pugui visualitzar en un mapa, sinó també aplicacions que puguin tenir en compte aquest component espacial per agrupar els resultats o per personalitzar-los depenent de la nostra ubicació. De la mateixa manera que les expressions temporals, es pot anar molt més enllà de la geolocalització d'entitats i detectar expressions geoespacials, fet que ens permet no només associar punts en un mapa sinó també rutes i direccions (Guerrero *et al.* 2011).

2.3. La polaritat i l'anàlisi del sentiment

Amb l'aparició del web 2.0 i l'increment del contingut generat pels usuaris al web (twitter, blogs, reviews, etc.) ha guanyat importància la idea de poder utilitzar el web per prendre el pols del que pensa la gent. És a dir, saber de què parla la gent (com ara els trending topics de twitter) i si en parla bé o malament. Hi ha una extensa literatura sobre els diferents treballs en aquesta àrea (Pang i Lee 2008) i diferents competicions (com RepLab 2012,[17] SemEval 2013)[18] que exploren diferent aspectes de l'anàlisi del sentiment i la polaritat.

Encara que aquest tipus de processament semàntic bàsic encara té molts problemes per resoldre, tant inherents a la complexitat del llenguatge (com ara la ironia) com d'altres inherents al mitjà (UGC). En certs tipus de documents (ressenyes, notícies, blogs) i mitjançant l'agregació de resultats, aquesta informació semàntica pot arribar a ser prou robusta com per permetre a les aplicacions una nova manera de presentar i buscar informació.

[15] http://www.simile-widgets.org/timeline
[16] http://developer.yahoo.com/boss/geo/
[17] http://www.limosine-project.eu/events/replab2012
[18] http://www.cs.york.ac.uk/semeval-2013/task2

3. APLICACIONS

Les següents seccions presentaran Correlator, Time Explorer i Yahoo! Quest. Aquestes aplicacions, desenvolupades a la Fundació Barcelona Media-Yahoo! Research Barcelona, proporcionen noves maneres de cercar i organitzar la informació basades en l'aplicació de tècniques de la lingüística computacional per obtenir automàticament els diferents tipus d'anotacions semàntiques descrites en els apartats anteriors.

3.1. Correlator

Correlator va ser la primera aplicació que el Laboratori de Yahoo! Barcelona va fer pública a yahoo! Sandbox (2008). Correlator feia cerques sobre la Viquipèdia anglesa, és a dir sobre unes dades socials (web 2.0) però uniformes i de caràcter neutre. Correlator va adreçar el problema de la cerca a gran escala (col·leccions grans de texts), l'aplicabilitat de les tecnologies de NLP a la cerca, així com l'escalabilitat d'aquestes tecnologies a les grans col·leccions de texts.

Correlator va ser innovador en el seu moment a nivell mundial, tot i que hi havia òbviament d'altres grups de recerca i nombrosos prototips que intentaven explorar altres paradigmes de cerca semàntica, fins i tot algunes que també feien servir entitats nombrades per visualitzar els resultats. La diferència d'aquests projectes respecte Correlator era la manera d'utilitzar aquestes entitats per ordenar els resultats, la granularitat d'aquests resultats (paràgrafs i entitats no documents) i la seva escalabilitat, ja que la majoria de la cerca en aquest camp es du a terme sobre col·leccions petites, dominis més tancats, degut al cost (tant en temps d' indexació com de consulta). Correlator permetia fer aquest tipus de cerca a gran escala (milions de documents).

FIGURA 1. CORRELATOR CREANT UNA PÀGINA SINTÈTICA PER A LA PREGUNTA "CLIMATE CHANGE KYOTO"

Per una banda, Correlator proporcionava una nova manera de fer cerques i d'accedir a la informació més enllà del paradigma de cerca per paraules (per exemple creant pàgines sintètiques amb trossos de les pàgines rellevants, veure Figura 1)

Correlator extreia i organitzava automàticament informació mitjançant tècniques automàtiques de processament de llenguatge i permetia fer cerques de caràcter més semàntic, per exemple per noms relacionats, per conceptes relacionats, per llocs o events relacionats. Els resultats de la cerca a Correlator no eren uns quants documents, sinó paràgrafs i entitats relacionades amb la nostra cerca. Correlator permetia a més visualitzar i fusionar tota aquesta informació, depenent del tipus de resultat (documents sintètics, mapes, línies de temps).

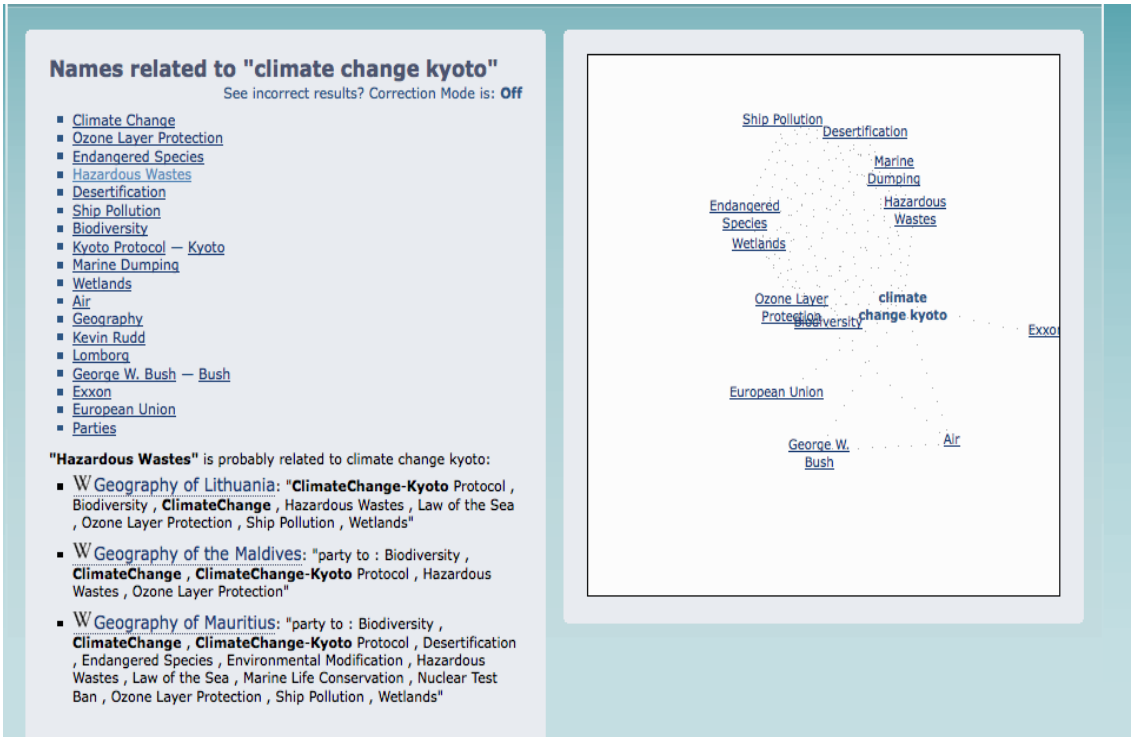

FIGURA 2. CORRELATOR ENTITATS NOMBRADES RELACIONADES AMB "CLIMATE CHANGE KYOTO"

La no sempre senzilla tasca de reconèixer i classificar les entitats rellevants per una cerca, pot permetre tant la millora dels resultats (trobem millor informació, nova informació, etc.) com la millora de l'experiència de l'usuari (noves maneres de visualitzar la informació, facilitar l'accés).

El paradigma de Correlator és més proper a la cerca de respostes ('Question Answering'), obtenint paràgrafs i entitats. L'obtenció d'aquestes entitats rellevants (veure Figura 2) no només és interessant per ella mateixa, sinó que ens permeten accedir i visualitzar la informació d'una manera diferent i en última instància millorar la pròpia cerca.

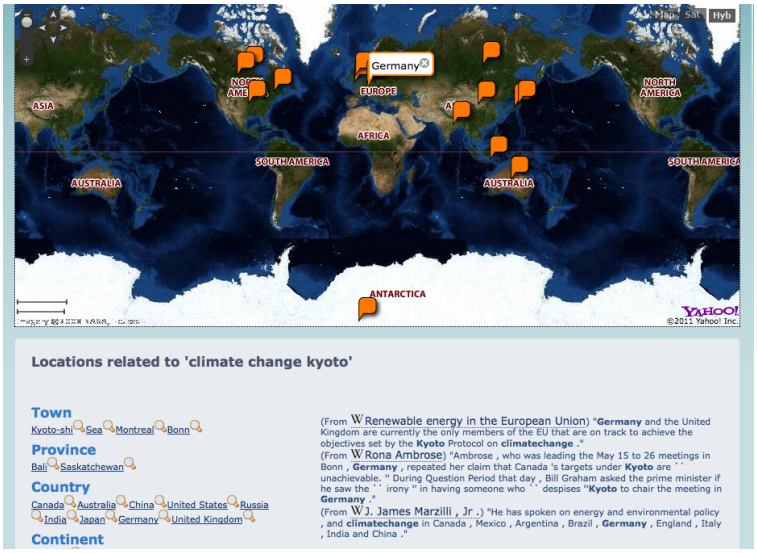

FIGURA 3. CORRELATOR GEOLOCALITZACIÓ D'ENTITATS EN UN MAPA

Correlator també permetia explorar d'altres dimensions: com l'espacial (veure Figura 3) permetent visualitzar les entitats geolocalitzades en un mapa; o la temporal (veure Figura 4) permetent visualitzar els resultats en ordre cronològic. I en tots dos casos, també els paràgrafs que justificaven la rellevància (i molts cops la relació) d'aquestes dates i llocs amb la cerca.

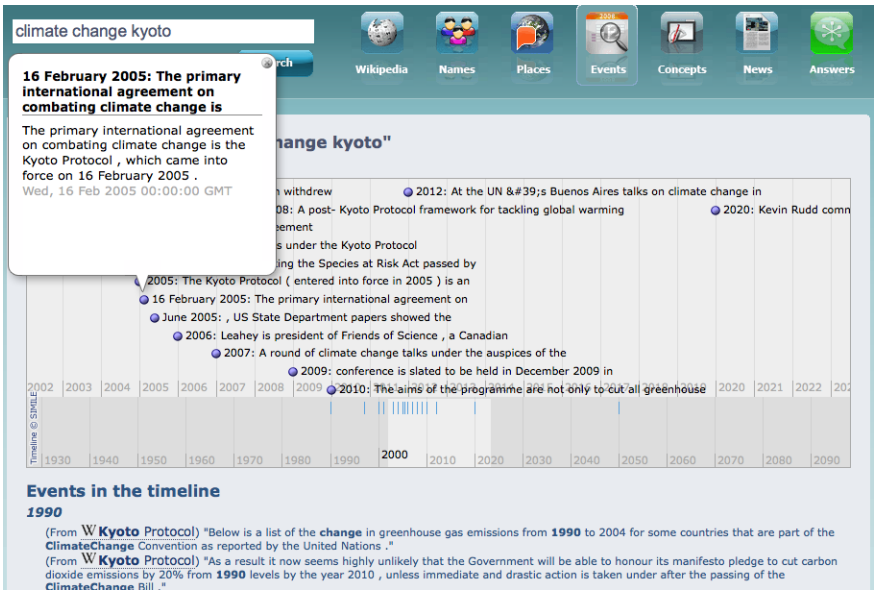

FIGURA 4. CORRELATOR VISUALITZACIÓ EN LÍNIES TEMPORALS

Aquestes diferents maneres d'estructurar i visualitzar la informació són les que permetien a l'usuari no només trobar i accedir molt més fàcilment a la informació, sinó també accedir a molta informació concreta i rellevant a la seva cerca, encara que estigués distribuïda en múltiples pàgines de la Viquipèdia anglesa i sense haver de visualitzar tots els documents sencers.

3.2. Time Explorer

La primera versió de Time Explorer[19] va veure la llum l'any 2010 dins del projecte Europeu Living Knowledge[20] i va ser escollit "People's Choice award at the 2010 HCIR Challenge" (Matthews *et al.* 2010). La principal novetat de Time Explorer és que anava més enllà que Correlator, no només en el gènere dels documents (blogs, notícies) sinó també en l'ús del component temporal.

El component temporal en els motors de cerca normalment s'ha vist relegat a intentar que els documents més recents apareguin com a més rellevants en els resultats de la cerca. En els últims anys però hi ha hagut un creixent interès a explotar no només la data de publicació del propi document (no sempre disponible) sinó també les possibles referències temporals contingudes en els propis documents (Baeza-Yates 2005). En alguns dominis, com per exemple les notícies, el component temporal és imprescindible per poder entendre la informació. Time Explorer permet poder explorar els resultats de la cerca sobre notícies, blogs, etc. en diferents components. Alguns d'aquests components provenen de metadades (autor, el tipus de document) però la majoria provenen del processament del propi contingut textual dels documents.

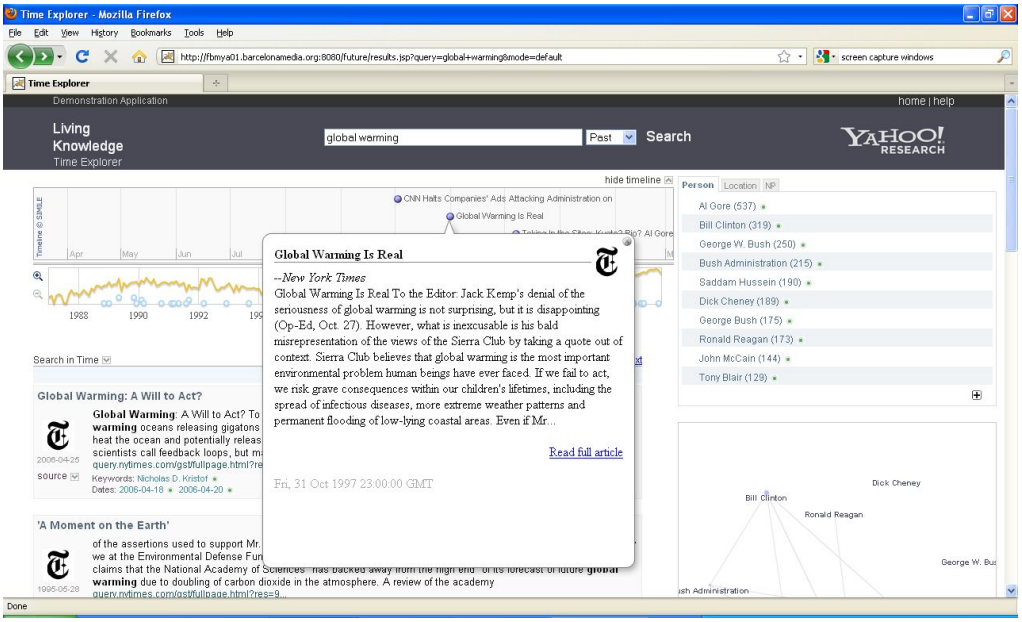

FIGURA 5. TIME EXPLORER

Time Explorer encara té moltes limitacions pel que fa a la comprensió dels documents, no només en la millora de la qualitat i cobertura de la informació extreta, sinó també en altres aspectes relatius a la comprensió del text: la detecció d'esdeveniments i els seus participants, la factualitat/negació, etc.

[19] http://fbmya01.barcelonamedia.org:8080/future/
[20] http://livingknowledge-project.eu

TimeExplorer deixa molta part d'aquesta interpretació de fragments de text recuperats (grups-paràgrafs) a l'usuari, que és qui realment acaba interpretant la informació recuperada.

3.3. YAHOO! QUEST

Un exemple completament diferent de les possibilitats de noves aplicacions que permet la lingüística computacional és Yahoo! Quest: una aplicació per ajudar els usuaris a trobar informació en una col·lecció de preguntes-resposta (en aquest cas Yahoo! Answers). Amb més de 21 milions d'usuaris diferents als Estat Units i 90 milions a tot el món Yahoo! Answers és un dels portals més grans on els usuaris comparteixen coneixement a la web en forma de preguntes i respostes.

Els actuals motors de cerca no són gaire útils per explorar aquest tipus de repositoris. El fet de desconèixer la resposta, fa que sigui més complicat fer una formulació exacta de la pregunta. Encara que molts cops altres usuaris han fet preguntes semblants és complicat trobar la resposta si la nostra pregunta no està formulada exactament igual. La Taula 1 mostra diferents preguntes en anglès sobre com evitar que un gos (*dog*) bordi (*bark*).

HOW DO I GET MY DOG TO STOP BARKING AT BIG DOGS?
How can i get my brother's dog to stop barking?
How can I stop my dog from barking all the time?
How can i stop my dog from barking at people?
How do i stop my dog from barking all the time?
How do you get the neighbors dog to stop barking?

TAULA 1. EXEMPLES DE DIFERENTS PREGUNTES SOBRE COM FER QUE ELS GOSSOS NO BORDIN

Yahoo! Quest es basa en un pre-processat del text per extreure informació dels documents per després poder-la indexar (veure Figura 6). De manera que quan l'usuari fa una cerca, es pugui recuperar aquesta informació i permeti a l'aplicació trobar preguntes semblants (i per tant respostes rellevants) i suggerir termes útils per refinar o reformular la pregunta de l'usuari. Els motors de cerca utilitzen índexs invertits (inverted indexes) per recuperar els documents rellevants per a una pregunta. Per poder recuperar i comptar les freqüències dels termes en els documents retornats com a resultat de la nostra pregunta de manera eficient, és necessari un nou tipus d'estructura que permeti recuperar els termes ràpidament (molt sovint anomenada forward índex). Quest utilitza un tipus especial de forward índex de MG4J.[21]

[21] http://mg4j.di.unimi.it/

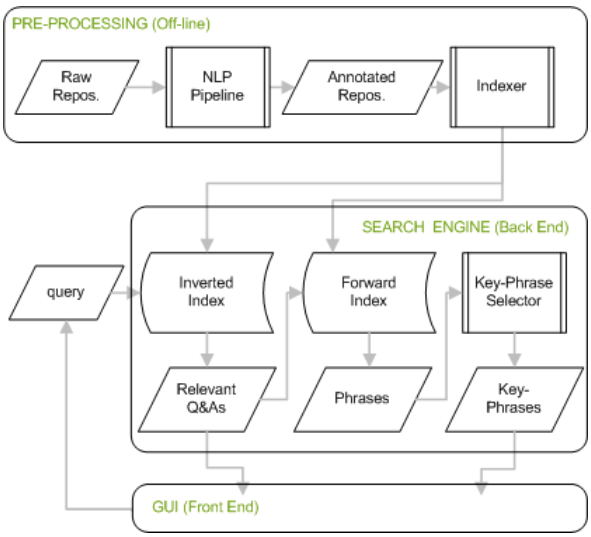

FIGURA 6. ARQUITECTURA DE YAHOO! QUEST

YAHOO! QUEST és una nova manera d'explorar i cercar preguntes i respostes de manera intuïtiva. Els resultats de la cerca s'augmenten amb un resum dels termes més importants entre els resultats, permetent refinar i explorar les respostes. A diferència d'altres mètodes (com ara l'expansió de preguntes) aquests termes no estan basats en la cerques d'altres usuaris, sinó que depenen exclusivament del conjunt de documents resultants.

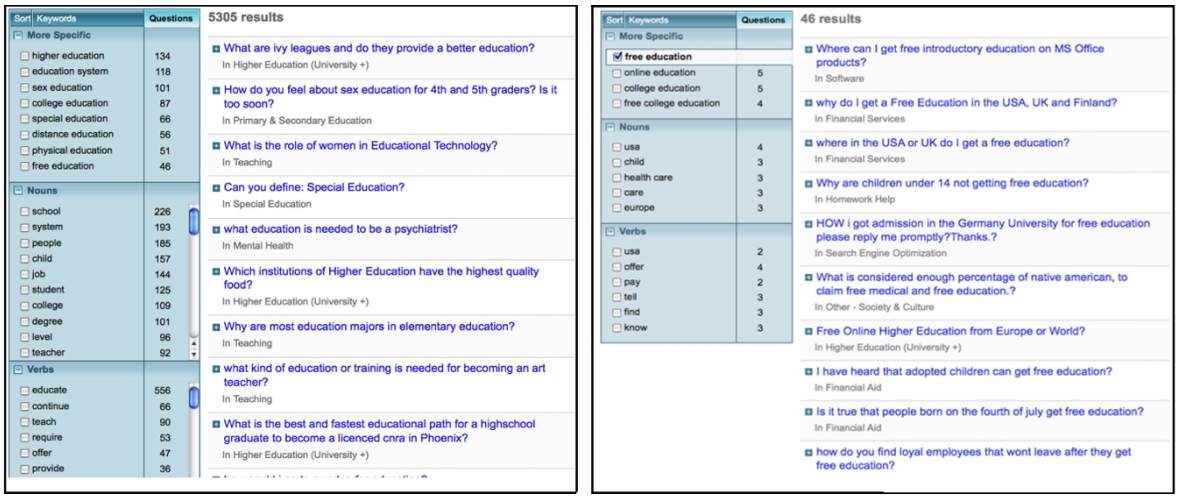

FIGURA 7. (ESQUERRA) RESULTATS PER "EDUCATION" (EDUCACIÓ) AMB DIFERENTS TERMES CLAU AGRUPATS PER POS. (ESQUERRA) ELS RESULTATS SELECCIONANT "FREE EDUCATION" (EDUCACIÓ GRATUÏTA).

Per extreure les expressions o grups sintàctics clau, formats per un o més termes rellevants d'un document, Yahoo! Quest fa servir l'arbre de dependències sintàctic obtingut de manera automàtica mitjançant Desr (Attardi i Ciaramita 2007). Per extreure aquest tipus de sintagmes clau, l'aproximació habitual és considerar els n-grames (n termes consecutius), amb l'explosió combinatòria que això representa de possibles sintagmes clau.

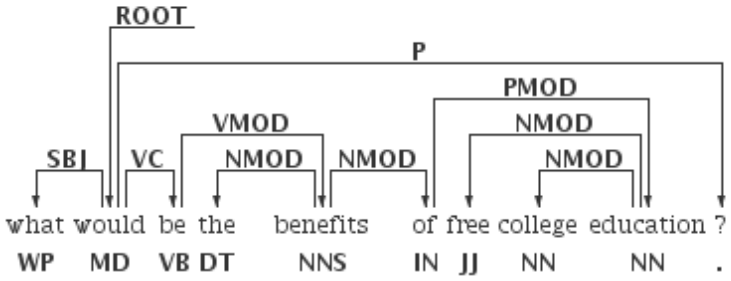

FIGURA 8. EXEMPLE D'UN ARBRE DE DEPENDÈNCIES SINTÀCTIQUES

L'extracció de grups/termes clau a Quest es du a terme en dos passos. El primer consisteix a extreure, per a cada nucli, el grup de paraules que governa. En l'exemple de la Figura 8, per exemple, a partir de "education" (educació) extrauríem el grup clau "free college education" (educació universitària gratuïta). El segon pas és extreure combinacions nucli-target que compleixin unes certes restriccions formals (descartant parelles con "what would") i d'interès (descartant "the benefits"). Aquestes restriccions es basen en el PoS d'ambdues paraules i el tipus de dependència sintàctica. A més, analitzant aquesta mateixa informació podem assignar heurísticament una categoria (Verbal, Nominal, Adverbial, etc.) als grups extrets.

Quest utilitza l'estructura de dependències sintàctica no només per restringir en quins n-grames estem interessats, sinó també permetent la construcció de d'expressions a partir de termes no correlatius en el text. Per exemple per extreure "free education" (educació gratuïta) a partir del text "free college education" (educació universitària gratuïta) encara que els termes "free" (gratuïta) i "education" (educació) no estiguin junts.

TIPUS	TERMES	TERMES ÚNICS
V	**11.850.760**	**691.406**
N	**12.364.631**	**3.583.684**

TAULA 2. NGRAMES EXTRETS PER 4.5 MILLONS DE PREGUNTES CORRESPONENTS LES CATEGORIES NOMINAL O VERBAL

La Taula 2 mostra el nombre de grups extrets sobre 4M preguntes. Yahoo! Quest[22] està disponible a Yahoo sandbox des del Novembre de 2009 sobre un conjunt de 4M de preguntes-respostes del servei de Yahoo! Answers en anglès. A l'aplicació final només es presentaven les classes Nominal i Verbal (24M grups) ja que la resta de classes semblaven no ser útils per als usuaris. També es va decidir separar en una llista especial els grups clau que contenien termes de

[22] Yahoo! Answers Comprehensive Question and Answers v. 1.0}, disponible a traves de Yahoo! WebScope http://webscope.sandbox.yahoo.com

la pregunta per tal de permetre a l'usuari poder refinar la seva pregunta ràpidament.

4. CONCLUSIONS

Els avenços en la lingüística computacional, com també el creixement de la informació estructurada i els nous dispositius mòbils (tauletes, telèfons intel·ligents), estan creant noves necessitats i noves aplicacions en la cerca d'informació.

La lingüística computacional encara és lluny de permetre una comprensió semàntica del contingut dels documents, però certes representacions semàntiques de parts del document ens permeten crear aplicacions que van molt més enllà de la cerca tradicional basada en cadenes de caràcters. Correlator, Time Explorer i Yahoo! Quest són alguns exemples d'aquests tipus d'aplicacions desenvolupades a la Fundació Barcelona Media-Yahoo! Research Barcelona.

BIBLIOGRAFIA

ATTARDI, G. i CIARAMITA, M. (2007), "Tree Revision Learning for Dependency Parsing", dins *Proceedings of the Human Language Technology Conference,* 388–395 [Consulta: 22 septembre 2013]. Disponible a: <URL *http://acl.ldc.upenn.edu/N/N07/N07-1049.pdf*>

BAEZA-YATES, R. (2005), "Searching the future", dins SIGIR Workshop MF/IR.

BLANCO, R. i ZARAGOZA, H. (2010), "Finding Support Sentences for Entities", dins *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR), 2010:339-346 [Consulta: 22 septembre 2013]. Disponible a: <URL http://www.hugo-zaragoza.net/academic/pdf/blanco_SIGIR2010.pdf> (DOI:10.1145/1835449.1835507)

DEMARTINI, G., MISSEN, M.M.S., BLANCO, R. i ZARAGOZA, H. (2010), "TAER: Time-Aware Entity Retrieval", dins *19th ACM International Conference on Information and Knowledge Management* (CIKM) pp. 1517-1520 (short-paper) [Consulta: 22 septembre 2013]. Disponible a: <URL http://www.hugo-zaragoza.net/academic/pdf/demartini_CIKM2010.pdf> (DOI: 10.1145/1871437.1871661)

GUERRERO NIETO, M., SAURÍ, R. i BERNABÉ POVEDA, M. A. (2011), "ModeS TimeBank: A Modern Spanish TimeBank Corpus", dins Revista de la SEPLN 2011.

GUERRERO NIETO, M. i SAURÍ, R. (to appear), "Porting Natural Language Time Expressions into Temporal Databases. A Corpus-based Methodology", dins *Spatial Humanities,* Crespo, A. (ed.), Cambridge Scholar Publishing.

MATTHEWS, M., TOLCHINSKY, P., BLANCO, R., ATSERIAS, J., MIKA, P. i ZARAGOZA, H. (2010), "Searching through time in the New York Times", dins *Bridging Human-Computer Interaction and Information Retrieval* (HCIR Workshop) [Consulta: 22 septembre 2013]. Disponible a: <URL http://www.hugo-zaragoza.net/academic/pdf/matthews_ HCIR2010.pdf>.

MIKA, P., CIARAMITA, M., ZARAGOZA, H. i ATSERIAS, J. (2008), "Learning to Tag and Tagging to Learn: A Case Study on Wikipedia", dins *IEEE Intelligent Systems* 23, 5 (September 2008), 26-33. [Consulta: 22 septembre 2013]. Disponible a: <URL http://www.hugo-zaragoza.net/academic/pdf/mika_ieee08.pdf>.

MOENS, M. F., JEUNIAUX, P., ANGHELUTA, R. i MITRA, R. (2006), "Measuring Aboutness of an Entity in a Text", dins *Workshop on TextGraphs*, at HLT-NAACL 2006, 25-28, New York City.

NADEAU, D. i SEKINE, S. (2007), "A survey of named entity recognition and classification", Journal of Linguisticae Investigationes 30:1, 3-26 (DOI: 10.1075/li.30.1.03nad)

NATTIYA, K., BERBERICH, K. i NØRVÅG, K. (2012), "Learning to Select a Time-aware Retrieval Model" SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 1099-1100. [Consulta: 22 septembre 2013]. Disponible a <URL http://www.l3s.de/~kanhabua/papers/SIGIR2012_L2S_TRank.pdf>

LREC workshop 2012 "@NLP can u tag #user_generated_content ?! via lrec-conf.org" URL http://nlp4ugc.barcelonamedia.org.

LLORENS, H., NAVARRO, B. i SAQUETE, E. (2009), "TimeML Temporal Expressions Detection for Catalan Using Semantic Roles and Semantic Networks", SEPLN, 43 (2009), pp. 13-21. [Consulta: 22 septembre 2013]. Disponible a <URL http://www.sepln.org/revistaSEPLN/revista/43/articulos/art2.pdf>

PANG, B. i LEE, L. (2008), Foundations and Trends in Information Retrieval 2(1-2), 1-135. (DOI: 10.1561/1500000011)

PUSTEJOVSKY, J., PATRIK, H., SAURI, R., SEE, A., GAIZAUSKAS, R. J., SETZER, A., RADEV, D. R., SUNDHEIM, B., DAY, D., FERRO, L. i LAZO, M. (2003), "The TIMEBANK Corpus", *Corpus Linguistics 2003: 647-656*

PUSTEJOVSKY, J., KNIPPEN, R., LITTMAN, J. i SAURÍ, R. (2005), "Temporal and Event Information in Natural Language Text", dins *Language Resources and Evaluation*, Volume 39, Issue 2-3, 123-164.

STRÖTGEN, J. i GERTZ, M. (2012), "Multilingual and Cross-domain Temporal Tagging", dins *Language Resources and Evaluation*, 2012, pp 3746-3753 Springer. (DOI: 10.1007/s10579-012-9179-y) [Consulta: 22 septembre 2013]. Disponible a: <URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/425_Paper.pdf >

ZARAGOZA, H., RODE , H., MIKA, P., ATSERIAS, J., CIARAMITA, M. i ATTARDI, G. (2007), "Ranking Very Many Typed Entities on Wikipedia" CIKM'07, [Consulta: 22 septembre 2013]. Disponible a <URL http://www.hugo- zaragoza.net/academic /pdf/zaragoza_CIKM07.pdf>.

AGRAÏMENTS

Els treballs i les aplicacions descrites en aquest article han estat finançades parcialment pels projectes Holopedia (MEC TIN2010-21128-C02-02) i Living Knowledge (ICT/2007.8.6) i han estat possible gràcies, entre d'altres, a Peter Mika, Roi Blanco, Michel Matthews, Pancho Tolchinsky, Mihail Surdeanu, Massimiliano Ciaramita, Giuseppe Attardi, Sebastiano Vigna, Paolo Boldi. També volem agrair l'ajut d'Elisabet Comelles en l'article.