Digit·HVM revista digital d'humanitats

Oportunitats per a l'enginyeria lingüística

Piek Vossen

Director tècnic en cap d'Irion Technologies
Vossen@irion.nl

Resum: Internet és tot comunicació i informació. No és un mer proveïdor de més i més dades, Internet, sinó que també implica una vertadera necessitat d'informació. Per dir-ho més clarament, hi ha una necessitat creixent de digerir quantitats immenses d'informació, la major part emmagatzemada en format text. En aquest article s'expliquen les oportunitats que ofereix l'enginyeria lingüística per desenvolupar productes d'alta qualitat aplicables a la societat de la informació. Es proposa un model general que serveixi per tractar les diferents etapes de l'anàlisi, l'emmagatzematge i l'accés a la informació i s'analitzen cada una d'aquestes etapes tot fent referència als sistemes principals i més utilitzats avui en dia. No es tracta d'una llista exhaustiva de totes les possibilitats i solucions, ni està completament actualitzada ni pretén estar-ho. La tecnologia es desenvolupa ràpidament i qualsevol intent per cobrir un camp determinat ja és directament obsolet a l'hora de publicar-lo. En comptes de tot això, però, l'autor vol oferir una perspectiva sobre els principals desenvolupaments actuals, i intentar encomanar al lector l'interès pel present i el futur de l'enginyeria lingüística.

1. Introducció

Fa deu anys no existia el terme enginyeria lingüística, i fins i tot avui en dia és un concepte completament nou per molta gent. Ben aviat, però, tothom en tindrà més que una mera idea: en un futur proper, la major part de la nostra tecnologia es manejarà a través de l'enginyeria lingüística i fins i tot s'hi basarà. Igual com el control remot i el ratolí són conceptes bàsics per a nosaltres, també ho seran els aparells de tractament de la parla i els assistents intel·ligents de lectura pacient. La meva filla petita amagava el comandament a distància per tenir el control de la televisió. La seva filla haurà d'ordir un pla diferent, ja que els nous aparells només faran cas de la seva veu, i en unes hores del dia limitades.

Aquesta diguem-ne apologia de l'enginyeria lingüística, deu anys abans, hauria provocat escepticisme, mofa i incredulitat. Que potser es tracta d'un altre intent desesperat d'atreure subvencions vers un projecte (sense sortida) per desenvolupar sistemes de comprensió del llenguatge?

Als anys vuitanta, molts inversors i polítics es van haver de refer després d'encetar projectes de traducció assistida a llarg termini, l'objectiu dels quals era traduir un grup d'oracions en un grapat de llengües diferents. Eren temps en què els lingüistes computacionals participaven tant en exercicis acadèmics com en formalismes sobre models lingüístics o teories sobre ordinadors, i amb prou feines es podia parlar d'un ús pràctic aplicable a aquests models, i ni tan sols de resultats indirectes. Amb el temps es van començar a desenvolupar algunes aplicacions, com ara els correctors ortogràfics i les eines automàtiques d'indexació i de resum, però la majoria no implicaven cap enginyeria lingüística. Molts pensaven (i encara ho fan) que el llenguatge humà és massa confús i complex, il·lògic, vague, ambigu i implícit per poder ser capturat dins un model.

Què ha canviat des d'aleshores? Sens dubte, tot excepte les llengües. En aquests deu anys hem assistit a una revolució silenciosa, una revolució resultant d'una altra de molt més sorollosa: Internet. A través d'Internet, els ordinadors poden accedir a una quantitat de text cada cop més gran en qualsevol idioma imaginable, i és amb Internet que els enginyers lingüístics, de sobte, han tingut accés a una quantitat mai vista de dades empíriques. Abans que aparegués Internet, els lingüistes esmerçaven anys de recerca per elaborar corpus lingüístics d'una mida mitjana, és a dir, textos seleccionats i tractats amb cura per tal de donar resposta a unes necessitats lingüístiques concretes.

Ara tenen a la seva disposició, a Internet, textos d'envergadura molt diversa, no només en anglès, sinó en la majoria d'idiomes del món.

Tanmateix, no són només les dades les que estableixen diferències: Internet és tot comunicació i informació. No és un mer proveïdor de més i més dades, Internet, sinó que també implica una vertadera necessitat d'informació. Per dir-ho més clarament, hi ha una necessitat creixent de digerir quantitats immenses d'informació, la major part emmagatzemada en format text. Si no pots trobar la informació rellevant a Internet, algú altre ho farà. Per alguns és més que una necessitat: és un problema d'informació.

De sobte sorgeix un mercat i una massa crítica de dades, i tothom pot jugar-hi i desenvolupar solucions. Les solucions, de fet, s'estan desenvolupant i gairebé a diari en trobem una de nova. Moltes són dolentes i la majoria amb prou feines recorren a l'enginyeria lingüística, però l'important és que permeten la creació d'un mercat amb unes bases que es poden aprofitar per desenvolupar altres aplicacions que utilitzen l'enginyeria lingüística per mostrar el seu ús. Un programa de recopilació estadística (*summarizer*) selecciona frases a partir de la freqüència amb què apareixen les paraules, cosa que proporciona una qualitat suficient segons el cas. Per tant, no seria gaire difícil afegir-hi anàlisis lingüístiques i millorar-ne el resultat, per exemple comptant les paraules lematitzades.

En aquest article espero poder explicar les oportunitats que ofereix l'enginyeria lingüística per desenvolupar productes d'alta qualitat aplicables a la societat de la informació. Així, en l'apartat següent vull proposar un model general que serveixi per tractar les diferents etapes de l'anàlisi, l'emmagatzematge i l'accés a la informació. En els apartats posteriors analitzaré cada una d'aquestes etapes: hi faré referència als sistemes principals i més utilitzats avui en dia, a solucions ràpides i econòmiques que s'avancen a la necessitat humana de tractar la informació o el coneixement d'una manera més convenient i, en la mesura que sigui possible, a sistemes d'enginyeria lingüística caracteritzats per un bon disseny i que es poden fer servir o desenvolupar per millorar i respondre a les expectatives. No es tracta d'una llista exhaustiva de totes les possibilitats i solucions, ni està completament actualitzada ni pretén estar-ho. La tecnologia es desenvolupa ràpidament i qualsevol intent per cobrir un camp determinat ja és directament obsolet a l'hora de publicar-lo. En comptes de tot això, però, voldria ser capaç d'oferir una perspectiva sobre els principals desenvolupaments actuals, i intentar encomanar al lector l'interès pel present i el futur de l'enginyeria lingüística.

2. Etapes d'anàlisi, emmagatzematge i d'accés a la informació

La manera més senzilla d'accedir a Internet és el WWW mateix, teclejant directament una adreça o seguint un enllaç, el qual et porta també a l'adreça que indica. Però Internet només dóna accés a informació, no l'assimila ni prova d'entendre-la. Això fa que la gent encara hagi de llegir-la (si és en format text) per comprovar-ne la rellevància. Com que Internet és tan gran, hi ha canvis constantment i els enllaços fan que et perdis de seguida, amb la qual cosa la gent que només navega es desespera fàcilment per trobar el que vol (llevat que sàpiga on és).

D'alguna manera, els sistemes informàtics poden ajudar els usuaris d'Internet assimilant part de la informació i proporcionant accés a aquesta informació prèviament digerida. S'hi poden aplicar diferents nivells d'anàlisi, que donen lloc a diferents representacions de la informació i que, en conseqüència, proporcionen maneres distintes d'accedir-hi i d'explotar-la. A la Figura 1 es pot observar una representació esquemàtica d'aquestes solucions. Al costat esquerre hi ha un grup de documents HTML que representen la informació a Internet. L'usuari pot accedir directament, a través d'un navegador, als documents individuals, un per un. A banda els HTML, hi ha per descomptat molts altres documents de text representats per DOC, PS o PDF i que no són accessibles amb un navegador. Baixant podem veure diferents maneres de compilar la mateixa informació i de donar accés de manera alternativa. Cap a la dreta trobem una primera fase d'anàlisi que deriva en una representació d'informació compilada, situada al mig. Apilats els uns a sobre dels altres, trobem diferents nivells de sofisticació de la informació compilada (índexs, jerarquies, fets i coneixement), que provenen de diversos processos d'anàlisi (indexació, classificació, extracció de dades i aprenentatge).

Al costat dret observem que les vies d'accés a la informació depenen de la sofisticació de l'anàlisi.

Figura 1. Etapes d'anàlisi

Els índexs són accessibles a través de la recuperació de paraules o motors de cerca. Teclejant unes paraules clau es poden obtenir els documents o pàgines HTML que més coincideixen amb aquestes paraules clau. El resultat és, doncs, una llista jeràrquica de documents o URL. Hi ha diferents maneres de muntar aquests índexs o d'analitzar les paraules clau, de combinar-les i expandir-les, i les tècniques lingüístiques es poden integrar fàcilment amb la tecnologia elemental del motor de cerca. Això es pot fer a una escala bàsica millorant l'anàlisi de la indexació i de la consulta, però també com un afegit en forma de *summarizers*, a més de millorar el reconeixement de llenguatges, el suport de diversos llenguatges, associant les consultes amb documents en altres llengües...

Un índex es pot considerar com una simple llista de termes normalitzats, un tipus de llista que pot significar el punt de partida per al desenvolupament d'una jerarquia o un arbre. Una jerarquia és una mena de classificació de dades o de documents per on es pot navegar com si fos un arbre, anant dels conceptes o les classes més generals als més específics (per exemple, de *Sports a Ball Sports* o *Water Sports*). A cada node de l'arbre hi podem trobar un grup de documents que estan relacionats amb el concepte. Hi ha diferents tipus de jerarquies, tal com veurem més endavant: tesaures, taxonomia, ontologia. Podríem dir que un tesaures es pot considerar com un raïm més global de paraules, documents o objectes pertanyents a alguna categoria (l'anomenada faceta). En canvi, les taxonomies i les ontologies són jerarquies de tots els objectes possibles i les seves propietats, definides més estrictament. L'enginyeria lingüística resulta útil per associar paraules i expressions a conceptes de l'ontologia, és a dir, trobar el significat correcte de les paraules a més de minimalitzar l'ontologia i relacionar automàticament els documents o termes als nodes de l'ontologia. Les ontologies i les classificacions ja són presents a Internet de moltes maneres, sobretot com a recursos estàtics (per exemple, les classificacions o els catàlegs de productes de Yahoo).

Una ontologia sempre captura relacions genèriques entre conceptes o classes, però no captura fets específics sobre exemples d'aquests conceptes. Una ontologia pot consignar que *una empresa té empleats*, però un fet consignarà que una empresa*concreta* té unes persones *concretes* com a empleats. Tot i que una ontologia pot ajudar en l'extracció de fets (defineix tots els fets possibles), no estipula quin és el cas en un cert moment del temps. Les ontologies són més persistents a través del temps (per exemple, *un estat pot tenir president*), mentre que els fets es concentren en un moment del temps determinat (*Clinton és el president dels EUA*). Tant els fets com les ontologies es poden extreure de dades textuals, però el procés d'extracció és molt diferent per cadascun. Les relacions ontològiques poden ser el resultat de l'anàlisi de grans quantitats de dades, en què es poden descobrir alguns models freqüents, mentre que un fet es pot expressar només un cop, i fins i tot en aquest cas pot no ser cert o antiquat.

El que hi ha de positiu als fets és que pots emmagatzemar-los en una base de dades relacional. Una

base de dades relacional és accessible mitjançant consultes SQL. Una consulta SQL consisteix en un comandament i unes referències a ítems en taules, per exemple EXPOSICIÓ+PRODUCTE (AUTOMÒBIL)+TIPUS (ASTROL)+TÉ (COIXÍ DE SEGURETAT). Una consulta SQL resulta complexa de formular i és aleshores que sembla escaient desenvolupar mòduls de Llenguatge Natural cap a SLQ que associïn consultes com ara *Té (vostè) una FURGONETA Astrol amb coixins de seguretat?*, amb consultes SQL. Les preguntes complexes es poden subdividir en d'altres de més senzilles, cosa que fa possible generar diàlegs planers en els quals pots consignar en primer lloc el que t'interessi i després especificar altres propietats i característiques. El comerç electrònic empeny amb força la representació de fets i les maneres en què s'hi pot accedir. Es tracta d'un petit pas des del catàleg de producte a la base de dades relacional. El comerç electrònic és una aventura mundial: no hi ha limitacions físiques perquè els possibles clients accedeixin a la teva zona. Això requereix que es puguin gestionar diàlegs o preguntes en diversos idiomes. A més, la qualitat del servei és més important que en els negocis tradicionals, ja que la competència tampoc no es veu limitada per fronteres físiques. La facilitat d'accés i de comunicació representa dues maneres bàsiques de distingir un negoci de la resta; un aspecte que vertebrarà el comerç del futur i, en conseqüència, el desenvolupament de l'enginyeria lingüística.

L'etapa final tracta les mateixes dades des d'una perspectiva diferent. En lloc d'interactuar directament amb els usuaris que volen accedir a la informació, l'usuari pot tenir un ajudant que operi en lloc seu. Amb la tecnologia agent entrem en una nova dimensió de l'accés a la informació: ara tenim un programari que intenta interpretar la informació. Resulta obvi que aquest programari pot accedir exactament als mateixos índexs, ontologies i fets que els humans (tot i que de manera més consistent i en quantitats més grans), però també cal dir que té molta menys capacitat per discernir què és útil del que no ho és. Els agents necessiten algun tipus d'intel·ligència per poder prendre decisions. Així, un agent o ajudant no només té accés a fets sinó que també ha d'adquirir coneixement. Per exemple, un usuari pot dir-li a l'ajudant que trobi el millor ordinador pel preu més baix; l'agent ha de generar un pla per reunir el coneixement suficient sobre la matèria i d'aquesta manera poder respondre la consulta amb el coneixement informàtic requerit (o, fins i tot, pot arribar a adquirir l'ordinador si és prou fiable). Si tota la informació s'emmagatzemés en format compilat, a un agent no li caldria l'enginyeria lingüística per aprendre. Tanmateix, com que la majoria de la informació està encara en format text, els agents necessiten ser capaços d'entendre tanta llengua com es requereixi. A més, els humans encara han de poder comunicar-se amb els agents, per la qual cosa l'enginyeria lingüística hi ha de ser present al cap i a la fi.

Hi ha encara una altra tendència que afectarà l'accessibilitat de la informació a Internet, i és que s'estan desenvolupant nous llenguatges d'etiquetatge, a banda d'HTML. L'XML (http://www.w3c.org/XML/) és un format més explícit que l'HTML. No només proporciona una representació comuna per a la composició dels documents, sinó que també ho fa amb relació al contingut. L'RDF (www.w3.org/RDF/), l'OIL (http://www.ontoknowledge.org/oil/) i el DAML (http://www.daml.org/) encara van més lluny per definir formalment el contingut mateix. L'RDF (*Resource Description Format*) integra diverses activitats de l'àmbit de les metadades al web, que inclouen mapes de llocs web, valoració de continguts, definicions *stream channel*, recopilació de dades amb motor de cerca (examinació de la Xarxa), grups de biblioteques digitals i creació distribuïda. L'RDF utilitza l'XML com a sintaxi d'intercanvi. L'OIL (*Ontology Interchange Language*) intenta combinar els models de la Xarxa amb representacions d'Estructura i Lògica de Descripció en enfocaments ontològics. L'OIL farà possible treure conclusions sobre el contingut representat en aquest llenguatge. El DAML (*Darpa Agent Markup Language*) és un formalisme adreçat a ajudar els agents de programari a interactuar entre ells. El llenguatge DAML també és una extensió de l'XML i l'RDF.

Cada un d'aquests estàndards és necessari per explotar la informació i els recursos als seus respectius nivells d'anàlisi. En part, això convertirà en obsolet l'esquema de difusió de la informació que hem vist més amunt, però al mateix temps demanarà eines que associïn automàticament text o parla amb aquestes representacions. En realitat, els formats en si no restitueixen el text a la fase d'anàlisi de la informació, sinó que es poden considerar com a formalismes de representació per a l'emmagatzematge de la informació analitzada, tal com es pot comprovar al centre de la Figura 1. Així, afectaran sens dubte les diferents maneres d'accedir a aquest coneixement i faran més fàcil desenvolupar programari per accedir a la informació compilada, ja que els desenvolupadors poden anticipar el format comú en què es representarà.

No continuaré analitzant els desenvolupaments d'aquests sistemes d'etiquetatge. Tampoc no entraré a parlar de panorames futuristes, en els quals trobem agents que fan prospeccions a la Xarxa per trobar coneixement i que formen comunitats per solucionar els problemes. En els apartats següents, em centraré sobretot en els sistemes d'indexació i cerca, de classificació i navegació i finalment de preguntes i respostes. Per a cada un, faré un cop d'ull a les pràctiques habituals avui en dia i n'analitzaré alguns exemples. De més a més, miraré d'apuntar les oportunitats que ofereix l'enginyeria lingüística integral (*build-in language technology*) i com millorar aquests sistemes.

3. Índex i cerca

Tothom coneix si fa no fa la primera generació de motors de cerca a Internet, com ara Yahoo (http://www.yahoo.com/) i Alta Vista (http://www.altavista.com/). Aquests motors indexen parts d'Internet i hi proporcionen accés mitjançant la cerca a través d'una paraula clau. L'objectiu d'aquests motors és cobrir la Xarxa i la realitat. Intenten donar accés a tantes pàgines web com poden i alhora miren d'actualitzar aquests enllaços amb regularitat.

És important de veure el que indexen realment i com associen les paraules clau als índexs. En la major part dels casos, els títols del web i les pàgines de l'índex s'utilitzen per muntar l'índex, cosa que no permet un accés directe al contingut de les pàgines web o a d'altres pàgines i lògicament tampoc no el permet a les que estan enllaçades a aquestes pàgines. A més, indexen cadenes i no tenen en compte la flexió, la funció gramatical ni l'estructura sintàctica. Per constatar les limitacions d'aquests motors de cerca, farem una ullada als exemples de consulta següents:

poisonous medication; poisonous medicine; poisonous medicines; toxic medication; toxic medicines; medicine for toxication; medicines for toxication; medicines against poisoning; medication for toxication; Help my kids took poison, show me medication; medicamento tóxico; medicamento intoxicación; medicina ponzoñoso; fármaco tóxico.

D'aquestes consultes, en podem extreure els punts següents:

1. Inclouen les formes del plural i del singular.

2. Inclouen consultes similars amb sinònims diferents.

3. Inclouen dues variants composicionals: una en què els medicaments són tòxics (1-5) i una altra en què es busca un medicament contra la intoxicació (6-9).

4. La consulta es pot fer en diferents llengües.

A partir d'un motor de cerca, caldria esperar el resultat següent:

1. No té en compte les variants flexives (p. ex., plural i singular) i dóna els mateixos resultats.

2. No té en compte l'ús dels sinònims i dóna els mateixos resultats.

3. Té en compte les diferències composicionals i mostra documents diferents per a cada interpretació.

4. Pot trobar la informació sense tenir en compte la llengua de la consulta.

Si observem els motors de cerca que hi ha a la Xarxa, comprovarem que cap funciona així. He inclòs una llista de resultats de cerca més avall perquè es puguin comparar els resultats. També es pot anar directament als llocs web i fer-lo directament des d'allà: s'hi pot comprovar aleshores que l'ús del singular o del plural, o d'un sinònim, dóna lloc a resultats molt diferents. Cap no és igual als altres. La indexació es basa en cadenes, i no hi té lloc cap normalització, tematització, anàlisi de compostos o anàlisi de derivats. A més, el significat composicional exacte no es té en compte en absolut, i casualment la mateixa paraula apareix com a ítem de l'índex per als mateixos documents, tot i que no sempre és aquest el cas. La relació entre els ítems no es té en compte en absolut:

WebSamples\Yahoo!_toxic_medication.htm
WebSamples\Yahoo!_poisonous_medicines.htm
WebSamples\Yahoo!_poisonous_medicine.htm
WebSamples\Yahoo!_poisonous_medication.htm
WebSamples\Yahoo!_medicine_for_toxication.htm
WebSamples\Yahoo!_medication_for_toxication.htm
WebSamples\Yahoo!_medication_against_poisoning.htm
WebSamples\AltaVista_toxic_medication.htm
WebSamples\AltaVista_poisonous_medicines.htm
WebSamples\AltaVista_poisonous_medicine.htm
WebSamples\AltaVista_medicine_for_toxication.htm
WebSamples\AltaVista_medicines_for_toxication.htm
WebSamples\AltaVista_medicines_against_poisoning.htm
WebSamples\AltaVista_medication_for_toxication.htm

És clar que, com que la indexació està basada en cadenes, una consulta en espanyol donarà documents en espanyol. Però, llevat que les paraules s'escriguin de la mateixa manera tant en anglès com en espanyol, no es poden obtenir documents en anglès amb una consulta en espanyol:

WebSamples\AltaVista_medicamento_toxico.htm
WebSamples\AltaVista_farmaco_toxico.htm
WebSamples\AltaVista_medicina_ponzoñoso.htm

Hi ha altres motors de cerca que intenten ser una mica més precisos pel que fa a la interpretació de la consulta. Més avall hi ha els resultats que, a partir de les mateixes consultes, es van obtenir d'Oingo i de Google. Oingo mira de presentar categories d'informació, però alhora ofereix l'opció de fer una cerca més estreta del significat dels termes consultats. Els significats provenen de la base de dades Wordnet (http://www.cogsci.princeton.edu/~wn/w3wn.html), una xarxa semanticolèxica de lliure accés. Conté un fons de conceptes amb relacions semàntiques entre si, a més de les associacions de paraules angleses a aquests conceptes. Els sinònims s'associen als mateixos conceptes i formen els anomenats synset (*synonymy set*). A Wordnet, *medicine* i *medication* són sinònims del mateix concepte, igual com*poisonous* i *toxic*. Podríem pensar aleshores que una expansió de les paraules consultades fins als sinònims corresponents implicaria un mateix resultat sense tenir en compte les paraules originals utilitzades en la consulta.

A la interfície d'Oingo s'han de seleccionar els significats de les paraules consultades manualment. Una vegada s'ha seleccionat el significat, Oingo pot trobar documents en els quals apareix la paraula consultada o un sinònim (per exemple, *toxic* en lloc de *poisonous*). Tal com podem comprovar en les pàgines obtingudes, els resultats no són tan espectaculars com es podia preveure. Les llistes resultants encara són molt diferents quan fem servir sinònims en les consultes:

WebSamples\Oingo_toxic_medicine.htm
WebSamples\Oingo_toxic_medication.htm

WebSamples\Oingo_poisonous_medicine.htm
WebSamples\Oingo_poisonous_medication.htm
WebSamples\Oingo_medicine_for_toxication.htm
WebSamples\Oingo_medicines_for_toxication.htm
WebSamples\Oingo_medication_for_toxication.htm
WebSamples\Oingo_medication_against_poisoning.htm

Pel que sembla, l'expansió als sinònims no resulta útil en tots els casos. SegonsVoorhees (1999), l'expansió de sinònims amb Wordnet pot tenir fins i tot un efecte negatiu sobre els resultats, sobretot si no se seleccionen els significats. No obstant això pot ser de gran ajuda, tal com es pot comprovar als exemples següents.

Hi ha una diferència essencial entre no seleccionar cap significat per a *organ* o seleccionar-ne un amb relació a *musical* (òrgan musical') o a *body part* (part del cos):

WebSamples\Oingo_organs.htm
WebSamples\Oingo_musical_organs.htm
WebSamples\Oingo_body_organs.htm

És una llàstima que s'hagin de seleccionar els significats a mà. No hi ha cap possible desambiguació, i no té gaire sentit desenvolupar un sistema de desambiguació d'aquestes característiques al lloc de consulta, ja que moltes consultes contenen una o dues paraules. Les consultes d'una o dues paraules no proporcionen context suficient per arribar a desambiguar.

Google no té en compte els significats diversos. En comptes d'això, llança una metacerca a diferents motors i hi aplica l'anàlisi del document per trobar els termes de consulta que són molt propers entre si. Igualment, mostra els fragments de text en què coapareixen les paraules; com que la memòria és immensa, encara pot generar molts més resultats.

WebSamples\Google_toxic_medicine.htm
WebSamples\Google_toxic_medication.htm
WebSamples\Google_poisonous_medicine.htm
WebSamples\Google_poisonous_medication.htm
WebSamples\Google_medicine_for_toxication.htm
WebSamples\Google_medicines_for_toxication.htm
WebSamples\Google_medication_for_toxication.htm
WebSamples\Google_medication_against_poisoning.htm

Tal com es pot comprovar, la limitació que suposa el fet que totes dues paraules han de coaparèixer pot conduir a un bon resultat. Sembla que no sempre cal seleccionar un significat concret. Així, Google explota l'alt grau de redundància que caracteritza la informació a Internet: aquesta s'emmagatzema diferents vegades i es formula en molts idiomes i de maneres molt diverses. El canvi que suposa l'emmagatzematge de la informació un sol cop en les mateixes paraules que les de la consulta és molt important. Més que escampar la consulta expandint-la en sinònims o altres expressions, sembla doncs més pràctic restringir-la a les coincidències literals solament. Òbviament, les coses canvien quan l'extracció s'aplica a petits grups de documents o intranets. En aquest cas, la informació pot ser expressada només un cop i en un sol document; aleshores l'expansió de la consulta resulta essencial per garantir-ne la recuperació.

Tant Google com Oingo intenten donar la impressió de precisió, però encara no posen gaire esment en el sistema de consultes. Així, no tenen en compte la variació fraseològica ni les relacions entre els termes de la consulta, amb la qual cosa resulta impossible tractar les diferències composicionals en el significat. Això no ens ha de sorprendre si ens adonem de les conseqüències d'una tal anàlisi. No només cal conèixer el llenguatge de cada document, sinó que també cal trobar el començament i el final de les frases (tokenització), analitzar gramaticalment les oracions per extreure'n les paraules lematitzades i les estructures composicionals, analitzar els compostos i derivats, detectar les

expressions multiparaula, descobrir relacions entre oracions creuades, determinar els significats de les paraules o les expressions, i d'altres. Tot això s'ha de fer per a cada llengua de treball. Els motors de cerca esmentats intenten abastar enormes parts d'Internet i necessiten actualitzar els seus índexs constantment. Una anàlisi lingüística dels documents i les consultes a aquesta escala demanaria un temps de processament enorme.

També hi ha proveïdors d'informació que procuren facilitar respostes més concretes. AskJeeves ha generat una expectació inusitada amb la il·lusió que podrien manegar vertaderes preguntes en llenguatge natural. Per desgràcia, val a dir que, a aquesta fita, no s'hi arriba a través de l'anàlisi i la comprensió de la pregunta, sinó a través d'una simple cerca de la pregunta en una base de dades on hi ha llistades totes les preguntes amb la resposta. Aquestes preguntes i respostes s'introdueixen manualment a la base de dades. Els resultats de la consulta que hem vist més amunt no són massa espectaculars, però segons com podem quedar-nos-en amb una bona impressió, tal com podem veure amb l'exemple *help1* de més avall. La consulta "Help my Kids took poison, show me the medication" (ajuda els meus fills han pres verí, mostra'm la medicació) té com a resultat, en realitat, la reformulació: "What should I do if my child swallows poison?" (Què hauria de fer si el meu fill s'empassés verí?).

WebSamples\AskJeeves_toxic_medicine.htm
WebSamples\AskJeeves_toxic_medication.htm
WebSamples\AskJeeves_poisonous_medication.htm
WebSamples\AskJeeves_medication_for_toxication.htm
WebSamples\AskJeeves_help1.htm
WebSamples\AskJeeves_help2.htm

No cal dir que aquest punt de vista és limitat. El nombre de preguntes i respostes és infinit i la informació emmagatzemada és difícil de mantenir i de controlar per als humans sense una ajuda addicional. Tan sols és qüestió de sort, el fet que el crit d'auxili coincideixi amb una pregunta prèviament emmagatzemada i que cobreixi el mateix contingut. Tal com es pot comprovar a *help2*, no sempre tindrem aquesta sort.

Resulta evident que tots els sistemes principals mostren una presència deficitària de l'enginyeria lingüística i que cap no té un caràcter "d'encreuament entre llengües" (*cross-linguistic*), és a dir, que pugui fer coincidir una consulta en espanyol amb documents en anglès. Ara per ara hi ha sistemes comercials que s'esforcen per millorar la tecnologia de cerca amb tècniques lingüístiques aplicades a molts idiomes i entre molts idiomes (Irion, Sail Labs, Textwise, Lexiquest). La major part d'aquestes solucions encara estan en fase de desenvolupament amb vista a petites intranets i dominis específics. La seva intenció és aconseguir una precisió més gran, o, en altres paraules, assolir l'objectiu que la resposta sigui entre els 10 primers resultats i que, si és possible, l'oració amb la resposta estigui subratllada al document. Aquests sistemes de recuperació de nova generació també maneguen diferents formes flexives i en alguns casos resolen compostos i expressions multiparaula. A més, el fet que s'apliquen sovint a grups de documents petits i homogenis dóna com a resultat una menor ambigüitat de significat. Per exemple, si els documents tracten de música, aleshores no cal desambiguar la consulta *d'organ*. La paraula només pot coincidir amb un significat de l'índex. Així, la recuperació d'alta precisió transmet la sensació de comprensió, però cal dir que en realitat aquests sistemes no entenen tampoc la pregunta. A part d'això, les diferències composicionals als exemples de consulta anteriors encara no es poden detectar. Ahttp://dis.tpd.tno.nl/21demomooi/ es pot comprovar el funcionament en directe d'un sistema de demostració que consisteix en una cerca multilingüe per a un grup concret de documents (sobre medi ambient a Europa). El sistema de recuperació TwentyOne, creat per TNO, preveu també les coincidències aproximades (*fuzzy-matching*), la qual cosa vol dir que els errors ortogràfics, els derivats i els compostos de la consulta poden coincidir amb els termes de l'índex. Per comparar, també es pot fer una ullada a Autonomy, els quals volen donar la imatge bastant explícita que són independents de les llengües i que no utilitzen l'enginyeria lingüística, mentre desenvolupen solucions per petites intranets i portals.

La recuperació interlingüística (*cross-lingual*) és factible normalment a través de diccionaris bilingües o d'una xarxa semàntica multilingüe. El projecte EuroWordNetva crear una xarxa d'aquest tipus per a 8 llengües: anglès, espanyol, italià, neerlandès, francès, alemany, txec i estonià, i s'hi afegeixen de

tant en tant d'altres idiomes. En el model d'EuroWordNet, els sinònims no només hi estan relacionats amb conceptes en cada idioma sinó també entre els idiomes via l'Índex Interlingual. Amb aquesta base de dades multilingüe *wordnet*, es pot aplicar una expansió a sinònims dins un mateix idioma (de *medicine* a *medication*) però també entre idiomes diferents (de *medicine* a *medicamento* i *medicina*). Les mateixes empreses treballen en aquests moments en la creació de recursos similars, i fins i tot els fan servir.

4. Classificar i navegar

Un dels desavantatges que tenen els motors de cerca és que mai no donen una idea clara del que realment existeix a la Xarxa. Una llista de resultats pot mostrar els documents que s'acosten a la consulta, però mai no se sap el que hi ha a banda d'això, i a més tampoc no es poden saber quins documents hi ha en el fons. Pel que fa a tot Internet això és possible perquè si fa no fa ho conté tot, però amb relació a petits grups de documents val la pena classificar la informació i presentar-la mitjançant arbres de categories. Yahoo va ser el primer gran motor de cerca que va utilitzar també categories que funcionen com a temes principals, dins dels quals es pot cercar més informació. Un altre exemple obvi és la versió electrònica de les*Pàgines Grogues* (http://www.yellowpages.com.au/). Les classificacions de Yahoo i les *Pàgines Grogues* s'elaboren manualment; la cobertura és necessàriament limitada i per tant no serveix per aclarir el que s'hi pot trobar.

Altres empreses creen sistemes que categoritzen automàticament documents.Adams (2001) estableix una distinció entre tres tecnologies de classificació:

1. Classificació per exemples: l'usuari elabora un conjunt de patrons representatius (*training set*) assignant manualment documents a categories. Els nous documents es classifiquen d'acord amb la seva similitud amb el*training set*. Empreses: Mohomine, Inxight, Autonomy.

2. Classificació estadística per extracció de paraula clau: es fan servir tècniques lingüístiques per extreure paraules clau i s'agrupen els documents que contenen paraules clau similars. Empreses: Semio, Cartia.

3. Basada en regles: regles explícites que capturen criteris a partir de quins documents es classifiquen com A o com B. Empreses: Verity.

En contrast amb la recuperació de documents per consulta, la classificació purament estadística i la classificació per exemples sembla que funcionen força bé sense la participació de l'enginyeria lingüística. Un document conté normalment text suficient per determinar la similitud amb un altre document. La variació resultant en paraules es pot constatar tot al llarg del document i les paraules generals i no concretes es poden deixar de banda ja que apareixen a tots els documents.

L'extracció de paraules clau resulta més pràctica amb una anàlisi lingüística, i algunes de les empreses esmentades recolzen força en l'extracció de les paraules clau més destacades amb vista a la classificació. En general, es pot afirmar que cada cop es necessiten més anàlisis lingüístiques com més petits són els documents. Per exemple, la classificació o filtratge de correu electrònic o URL resulta més difícil sense aplicar-hi una associació semàntica o lingüística. S'ha de reconèixer el tema a partir d'una sola línia temàtica. La classificació només és possible si els significats individuals estan relacionats amb dominis i aquests significats es poden seleccionar amb un mètode de desambiguació.

Un problema específic que sorgeix a l'hora de classificar documents és el mètode d'accés i visualització. Una manera habitual de visualitzar la classificació és l'arbre, però es tracta d'estructures que poden esdevenir massa grans i complexes, cosa que n'obstaculitza l'ús. Per fer-hi front de manera dinàmica, actualment es treballa en diverses solucions tecnològiques. Els enllaços següents

mostren alguns bons exemples dinàmics de tot plegat:

Reuters: http://reuters.medialab.nl/aqua.htm

WebBrain: http://www.webbrain.com/open_IE.htm

Inxight: http://www.inxight.com/products_wb/tree_studio/tree_studio_demos.html

Un desavantatge inherent a tota classificació és que obliga l'usuari a accedir a informació des d'un punt de vista concret. Si la classificació és gran i complexa, l'usuari s'hi pot perdre. Pot estar buscant una distinció errònia (que no estigui feta o que la informació desitjada estigui classificada de manera diferent), o buscant la distinció correcta en el lloc equivocat. Per solucionar això, o bé l'usuari ha de poder ser capaç d'organitzar la classificació d'acord amb el seus gustos, o bé es pot complementar la classificació amb una opció de cerca. En el primer cas, ha de ser possible extreure múltiples vistes d'associacions de classificacions i aleshores l'usuari en pot seleccionar una. L'estructura subjacent pot incloure múltiples classificacions dels mateixos documents i múltiples relacions entre classes. De manera alternativa, un usuari pot introduir-hi una classe, que es pot redireccionar a una categorització en ús. En aquest cas, hi ha un índex a part de paraules a categories.

Hi ha algunes iniciatives per desenvolupar representacions estandarditzades de la mateixa informació en maneres diverses. Les anomenades associacions de temes s'utilitzen per mostrar la mateixa informació des de qualsevol perspectiva. Se'n pot trobar més informació a: http://www.gca.org/papers/xmleurope2000/papers/s22-04.html. El programari de visualització es pot crear sobre la base d'aquest model.

Hem vist productes que classifiquen documents. La classificació d'un document no és realment una ontologia. Hi ha, però, altres maneres semblants d'estructurar la informació. Així, dins el comerç electrònic, moltes empreses posen a l'abast catàlegs dels seus productes, i aquests catàlegs també es poden considerar com una mena de classificació, tot i que no necessàriament associada amb documents. Ara, automatitzar l'elaboració de catàlegs és molt difícil, ja que, sovint, les descripcions dels productes són curtes i les categories no es deriven sempre de les descripcions. També cal tenir en compte que alguns catàlegs contenen milions de productes i la seva accessibilitat presenta molts problemes. A més, les empreses poden demanar que es determini la manera exacta com s'organitzarà la classificació. Comparats amb la informació en documents, els catàlegs són més pobres però alhora més sistemàtics: cobreixen normalment només uns pocs tipus de conceptes amb un nombre limitat de propietats o característiques. Una manera òbvia de tractar els catàlegs és convertir-los en bases de dades relacionals. Tot plegat ho analitzarem en l'apartat següent.

5. Extracció de dades i sistemes de pregunta-resposta

Un catàleg pot estar dissenyat d'acord amb una estructura jeràrquica, igual com una classificació, però les jerarquies, en el cas del catàleg, resulten menys complexes i intenses. El més interessant dels catàlegs són les característiques que defineixen els productes. Així, sovint podem observar als llocs web de comerç electrònic descripcions de les característiques (preus, data de lliurament, colors, mides, quantitat) i un nombre limitat d'opcions. Aquesta estructura permet ser emmagatzemada en una base de dades relacional, amb la qual cosa, un cop emmagatzemada, hi podem fer preguntes molt específiques sobre productes amb unes característiques concretes. Es tracta doncs d'una informació tant ontològica com factual, i són les limitacions ontològiques les que dictaminen les propietats o les característiques del producte o el tipus de producte. En darrer terme, aquest és el model que mostrarà la base de dades. Tot plegat, els mateixos productes (números de sèrie) i el seu

estatus (les propietats en si) es poden considerar com els fets que s'expressen a les taules d'una base de dades.

Moltes empreses desenvolupen en aquests moments sistemes per emmagatzemar "coneixement" sofisticat en bases de dades amb la idea de proporcionar accés a aquest coneixement. En tant que la informació hi és present en forma de documents, la informació general i el suport de productes s'hi poden aplicar a través de la indexació i la classificació, tal com s'ha descrit més amunt. Això, però, no condueix al coneixement específic: per tal d'obtenir un coneixement més detallat, hi ha empreses que emmagatzemen preguntes i respostes específiques en bases de dades. D'aquesta manera, ofereixen solucions a problemes concrets. Diferents tipus de coneixement es munten de diferents maneres. No els analitzaré tots, però sí que n'esmentaré exemples per donar-ne una perspectiva general.

La solució més senzilla és emmagatzemar o "enllaunar" preguntes i respostes de la manera que ho fa AskJeeves per a la informació general. Hi ha empreses que ho fan a través de l'emmagatzematge de problemes i solucions concrets ja coneguts per a un producte. Aquesta informació, de vegades, s'extrau manualment de documents i manuals, que en alguns casos es basen en el sistema PMF (preguntes més freqüents), i d'altres vegades es fa a través del registre de consultes de l'usuari i respostes, o mitjançant algun tipus de diàleg de diagnòstic que extregui coneixement de la resposta a consultes i que generi possibles preguntes relacionades amb aquell coneixement. Com que desenvolupen aquests sistemes per a clients concrets, poden perfectament crear sistemes de suport d'una alta sofisticació per integrar-los en, per exemple, centres d'atenció telefònica o serveis d'assistència tècnica.

Dos exemples d'aquesta mena d'empreses es poden trobar a:

ServiceWare: http://www.serviceware.com/

Demo: http://www.serviceware.com/solutions/essdemo.asp

Primus http://www.primus.com/

Demo: http://www.primus.com/search.asp

El fet que aquestes empreses es dediquin sobretot a tecnologia relacionada amb l'àmbit de preguntes i respostes dóna la impressió que automatitzen l'explotació de coneixement. Tanmateix, la seva tasca consisteix bàsicament en retallar els costos de les empreses amb una automatització intel·ligent d'alguns dels seus serveis de suport.

No cal dir que algunes d'aquestes empreses no confien necessàriament en l'enginyeria lingüística, però no s'esdevé el mateix amb les empreses que es dediquen a extreure coneixement a partir de dades estructurades (bases de dades) i no estructurades (text) proporcionades pels clients. El procés clau és l'extracció de la informació, la qual es basa en part en l'enginyeria lingüística i en part en el coneixement del domini. El coneixement de domini funciona com una ontologia que limita la possible informació cercada. L'enginyeria lingüística s'utilitza per extreure informació del text que coincideix amb aquesta ontologia. Així, el procés consisteix bàsicament en el farciment de plantilles, en què l'ontologia defineix les possibles plantilles i l'anàlisi textual en dóna com a resultat el farciment. Com que l'ontologia és petita i explícita, la part de comprensió del llenguatge pot extreure'n dades fiables. Només interpretarà expressions i frases que tinguin sentit dins el marc interpretatiu de l'ontologia, amb la qual cosa resultarà evident que les diferències composicionals, com en el cas de *poisonous medicine* (medicament tòxic) i *medicine for poisoning* (medicament contra la intoxicació), són essencials per a l'extracció d'informació.

Per ampliar dades sobre aquests sistemes d'extracció d'informació, podeu consultar Gaizauskas i

Humpreys (1997).

Com a exemples de sistemes comercials que es dediquen sobretot a l'extracció d'informació podem esmentar:

iPhrase: http://www.iphrase.com/

ClearForest http://www.clearforest.com/

Totes dues empreses utilitzen tècniques lingüístiques per interpretar el text i les frases amb vista a emplenar plantilles sobre productes i extreure ontologies. A la Figura 2 es pot comprovar l'arquitectura que utilitza ClearForest: es fa servir una definició dels conceptes i les relacions en forma de "reglament" per extreure el contingut del text, i els reglaments es munten prèviament per als dominis.

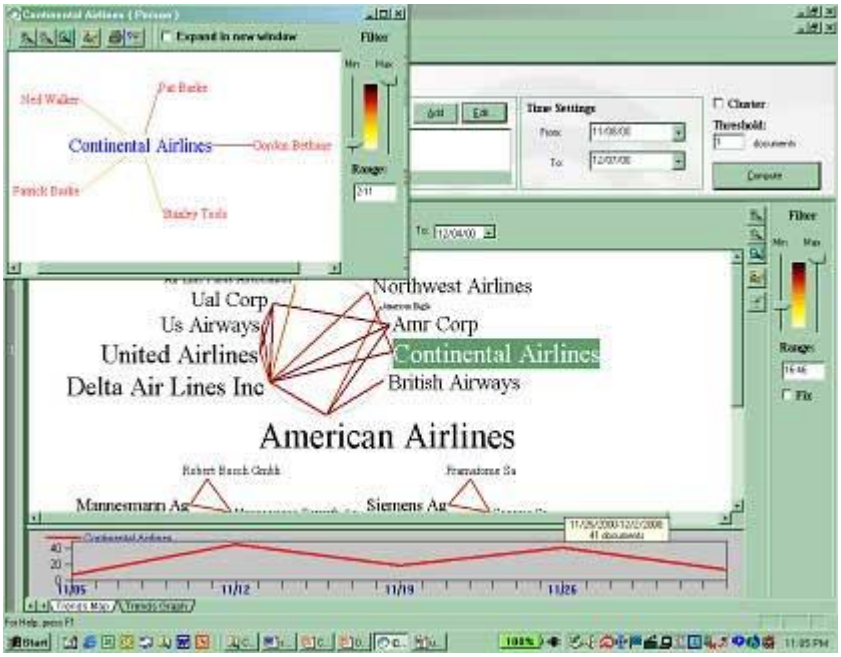

Figura 2. Relacions a ClearForest

Una ontologia sempre captura relacions genèriques entre conceptes o classes, però no captura fets específics sobre exemples d'aquests conceptes. Una ontologia pot consignar que una *empresa* té *empleats*, però un fet consignarà que una empresa*concreta* té unes persones *concretes* com a empleats. Tot i que una ontologia pot ajudar en l'extracció de fets (defineix tots els fets possibles), no estipula quin és el cas en un cert moment del temps. Les ontologies són més persistents a través del temps (per exemple, *un estat pot tenir president*), mentre que els fets es concentren en un moment del temps determinat (*Clinton és el president dels EUA*). Tant els fets com les ontologies es poden extreure de dades textuals, però el procés d'extracció és molt diferent per cadascun. Les relacions ontològiques poden ser el resultat de l'anàlisi de grans quantitats de dades, en què es poden descobrir alguns models freqüents, mentre que un fet es pot expressar només un cop, i fins i tot en aquest cas pot no ser cert o antiquat.

La Figura 3 mostra com s'extreuen taxonomies de documents concrets. En aquest exemple s'han extret noms de persona; per a cada persona es poden trobar i es poden expressar dades diferents.

Figura 3. Taxonomia ClearForest

Al lloc d'iPhrase es pot comprovar, mitjançant demostracions, la manera com aquest sistema proporciona accés a la informació: http://www.iphrase.com/demo. La seva anàlisi de dades permet tractar preguntes complexes i iteracions de preguntes com ara:

- Quines furgonetes tenen coixí de seguretat?
- Disposa l'Astro també d'un lector de CD?

També poden generar taules amb una perspectiva general que continguin preus i propietats, i presentar-les als clients que ho demanin. Un cop feta la primera pregunta, poden oferir una taula amb totes les *furgonetes* disponibles equipades amb*coixí de seguretat* i especificar altres dades com ara *marques* i *preus*. La segona pregunta s'interpreta aleshores dins el context que s'ha creat per a la primera. Gràcies a la rica base de dades de què disposa, iPhrase pot tractar la pregunta al mateix nivell que una consulta SQL.

EasyAsk és una empresa especialitzada justament en aquest àmbit. Han desenvolupat un complet sistema de comerç electrònic en el qual les bases de dades relacionals s'estenen amb un llenguatge natural a la interfície SQL. El sistema funciona perquè reconeix algunes paraules en la consulta com a ordres SQL i d'altres com a noms per taules. Una consulta com "*Mostra'm* totes les *furgonetes* amb *coixí de seguretat*?" es pot tractar perquè "Mostra'm" és l'ordre i "furgonetes" i "coixins de seguretat" són ítems que pertanyen a unes taules concretes. El sistema cercarà productes que tinguin relació amb els dos ítems de la taula i mostrarà una llista o una taula de resultats. Per tant, la consulta no demana gaire processament per arribar a una anàlisi de la consulta d'aquestes característiques. N'hi ha prou amb una senzilla llista d'ordres, noms de taula i alguns sinònims.

Hi ha una versió *demo* disponible a http://www.easyask.com/demo/. Mentre que iPhrase dóna més importància a l'extracció de dades i l'anàlisi lingüística de les preguntes i respostes, EasyAsk se centra més en una solució genèrica que es pugui aplicar a qualsevol base de dades relacional. L'avantatge que representa d'EasyAsk és que resulta senzill d'aplicar a qualsevol base de dades existent sense que calgui a penes personalitzar-la.

La Figura 4 mostra el disseny del sistema iPhrase. La base de coneixement de domini fa el mateix paper que el reglament de ClearForest. A més de la base de coneixement, iPhrase ofereix una

sofisticada interfície lingüística per analitzar les consultes i associar-les a la base de dades, a més d'un component de generació de respostes:

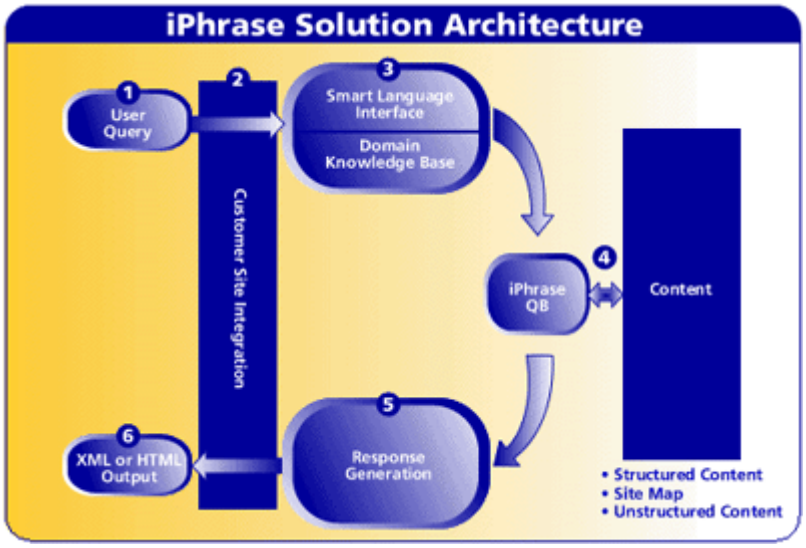

Figura 4. Arquitectura del Sistema iPhrase

La propera fase dels sistemes comercials podria ser el desenvolupament de sistemes de diàleg damunt la base de dades relacional. A començament dels vuitanta es van desenvolupar diversos sistemes de diàleg (comercials i experimentals), una bona visió global dels quals es pot trobar a Jönsson (1997). Òbviament, el diàleg demana uns models i unes tècniques lingüístiques més sofisticats, com ara:

- Entendre les preguntes al nivell d'un acte de parla per diferenciar entrepeticions, ordres, aclariments, etc.
- Analitzar les referències anafòriques contingudes en les preguntes, com araPuc comprar-lo?, on —lo fa referència a una entitat prèvia.
- Proporcionar una resposta sobre quines preguntes es poden respondre i quines no: Hi ha a prop una piscina?
- Proporcionar una resposta sobre per què una pregunta no ha rebut cap contesta: processament del llenguatge o adequació del contingut.
- Fer servir preguntes aclaridores de manera intel•ligent per resoldre ambigüitats o limitar la quantitat d'informació que es dóna: una llista de 200 hotels pot resultar excessiva.

El desenvolupament de bons sistemes de diàleg és difícil i delicat. L'ús dels sistemes que intenten imitar la mímica humana pot esdevenir fàcilment tediós, ja que la gent espera resultats i no vol perdre el temps amb una màquina que no entén les intencions ni els esforços comunicatius de l'usuari. Ara, si les bases de dades relacionals com les que acabem de veure s'estenen més i més en l'àmbit del comerç electrònic, sorgirà una necessitat cada vegada més gran d'accedir-hi amb uns sistemes de diàleg limitats. El sistema iPhrase ja hi treballa i aviat podrem gaudir de més sistemes similars.

6. Altres desenvolupaments

Hi ha dos interessants desenvolupaments pel que fa a la recuperació de la informació i els portals d'informació que estan relacionats directament amb l'enginyeria lingüística:

- La personalització de la informació
- L'autorització de la informació.

La personalització és un instrument relacionat amb la intel·ligència artificial que està concebut per dissenyar els perfils d'usuari. A partir d'aquests perfils, els usuaris poden rebre un servei més adequat només amb la informació que els interessa. El meu perfil pot informar que estic interessat en els instruments musicals i no en medicina. Aleshores, la consulta sobre *organ* que hem vist a l'apartat 3 es pot interpretar directament com una consulta sobre un òrgan musical. Amb els sistemes de classificacions passa el mateix, i només es mostraran les categories que interessin. Els perfils es poden muntar consignant de manera explícita alguns interessos, però també controlant l'activitat internauta de la persona, és a dir, què visita, què baixa o què llegeix. Un aspecte interessant dels perfils és que es poden agrupar i es poden utilitzar per desenvolupar o oferir serveis concrets a grups importants d'usuaris. També són interessants perquè les interaccions prèvies amb els clients a Internet es poden emmagatzemar en un historial personal, utilitzable per referir-s'hi en la comunicació.

L'autorització d'informació pren cada vegada més importància: com més informació hi ha disponible, més difícil resulta verificar-ne la qualitat. Un dels desenvolupaments naturals d'Internet, doncs, serà l'organització dels usuaris en petites comunitats (taulers d'anuncis, xats, llistes de correu), unes comunitats que poden concentrar la comunicació i la informació. L'avantatge resultant és que la comunitat o el grup pot controlar la informació intercanviada, un control que es podria veure com a procés d'autorització. Si una comunitat de programadors interessats en objectes determina que un programari gratuït concret és un bon sistema de base de dades, aleshores aquest programari adquirirà un cert valor. És possible avaluar l'ús de la informació en una comunitat o reconèixer els comentaris valoratius que se n'hi fa ("Aquest programa és excel·lent!", per exemple) i emmagatzemar aquesta informació. Aleshores es podrà afinar la recuperació d'informació amb l'objectiu d'obtenir la resposta més adient.

Bibliografia:

FELLBAUM, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

GAIZAUSKAS, R. y HUMPHREYS, K (1997). "Using a semantic network for information extraction". *Natural Language Egnineering*, vol. 3, part 3&3, p. 147-169.

JÖNSSON, A. (1997). "A model for habitable and efficient dialogue management for natural language interaction". *Natural Language Engineering*, vol. 3, part 3&3, p. 103-121.

VOSSEN, P. (ed.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Dordrecht: Kluwer Academic Publishers.

VOORHEES E. M. (1999). "Natural Language Processing and Information Retrieval". A: *Information Extraction: Towards Scalable, Adaptable Systems*. Springer (Germany): M. T. Pazienza (ed.), p. 32-48.

Enllaços relacionats:

- W3C, World Wide Web Consortium: XML
  http://www.w3.org/xml
- WordNet
  http://www.www.cogsci.princeton.edu/~wn/

- Euro WordNet
  http://www.hum.uva.nl/~ewn/
- Reuters Medialab: Aqua
  http://www.reuters.medialab.nl/aqua.htm

[Data de publicació: desembre de 2001]

Citació recomanada:

VOSSEN, Piek (2001). “Oportunitats per a l'enginyeria lingüística”. *Digithum*, núm. 3 [article en línia].
DOI: http://dx.doi.org/10.7238/d.v0i3.597