Cerca i recuperació d'informació

Antoni Oliver

PID_00159195

Universitat Oberta de Catalunya
www.uoc.edu

Índex

Introducció..... 5

Objectius..... 7

1. Recuperació d'informació..... 9
1.1. Components d'un sistema de recuperació d'informació 9
1.1.1. Preprocessament de documents 9
1.1.2. Construcció de l'índex del document 10
1.1.3. Processament de les consultes i cerca dels documents que satisfan la consulta 12
1.1.4. Avaluació de sistemes de recuperació d'informació 14
1.2. Tècniques de processament del llenguatge natural aplicades a la recuperació d'informació 15
1.2.1. Millora de la indexació fent servir informació morfològica 15
1.2.2. Millora de la indexació fent servir informació sintàctica 19
1.2.3. Millora de la indexació fent servir informació semàntica 19
1.3. GATE com a sistema de recuperació d'informació 21
1.3.1. Instal·lació i execució de GATE 21
1.3.2. Activació del connector de recuperació d'informació ... 22
1.3.3. Creació del corpus 22
1.3.4. Creació del *datastore* i assignació del corpus al *datastore* 23
1.3.5. Indexació del corpus 24
1.3.6. Creació d'un recurs de processament per a la cerca 24
1.3.7. Creació del *pipeline* application 25
1.3.8. Realització de la cerca 25
1.4. Conclusions 26

2. Extracció d'informació..... 28
2.1. Reconeixement d'entitats amb nom 29
2.1.1. Estratègies per al reconeixement d'entitats amb nom ... 32
2.2. Resolució de la coreferència 36
2.3. Extracció d'esdeveniments 36
2.4. GATE com a sistema d'extracció d'informació 38
2.4.1. Detecció d'entitats amb nom amb GATE 39

3. Sistemes de pregunta-resposta..... 40
3.1. Arquitectura dels sistemes de QA 41
3.1.1. Mòdul de processament de les preguntes 41

3.1.2. Mòdul de processament dels documents 41
3.1.3. Mòdul d'extracció la resposta 42
3.2. Conclusions 42
3.3. El sistema QA START 42
3.3.1. Funcionament bàsic 42

4. Recuperació d'informació multilingüe..... 44
4.1. Aproximacions per a la recuperació d'informació multilingüe ... 44
4.1.1. Ús d'un vocabulari controlat per a la indexació i la recuperació 44
4.1.2. Ús de sistemes de traducció automàtica 45
4.1.3. Ús del model d'espai vectorial de Salton 45
4.1.4. Ús de corpus multilingües 45
4.2. Extracció d'informació multilingüe 46
4.3. Sistemes de pregunta-resposta multilingües 46

5. Els cercadors d'Internet..... 47
5.1. *Open Directory Project*..... 48
5.2. L'algorisme Page Rank™ de Google 48
5.3. Funcionalitat de cerca avançada de Google 50
5.4. Incloure la meva pàgina web en els cercadors 51
5.5. Alertes de Google 51
5.6. Google Academics 51
5.7. Google Books 52
5.8. SandBox de Yahoo 52
5.9. Altres serveis oferts pels principals cercadors 53
5.10. Finançament del cercadors d'Internet 53
5.11. Conclusions 54

Resum..... 55

Bibliografia..... 57

Introducció

En els darrers anys la quantitat d'informació disponible de manera immediata ha augmentat espectacularment. Abans de la popularització d'Internet calia desplaçar-se a centres específics (biblioteques, arxius, etc.) per a poder accedir a cert tipus d'informació. Amb Internet tenim accés a tot això i a molt més des de casa o des del nostre lloc de treball. Aquesta quantitat enorme d'informació de res no serviria si no es pot cercar i recuperar d'una manera eficient.

En aquest mòdul estudiarem diverses tasques relacionades amb l'accés i l'extracció de la informació continguda en documents textuals. Veurem que moltes d'aquestes tècniques funcionen força bé sense fer un ús intensiu de tècniques de processament del llenguatge, però en molts casos aquestes tècniques poden ajudar a millorar els resultats. En aquest mòdul estudiarem les tasques següents:

- Recuperació d'informació, en anglès *information retrieval* (IR)
- Extracció d'informació, en anglès *information extraction* (IE)
- Sistemes de pregunta-resposta, en anglès *question answering* (QA)
- Recuperació d'informació multilingüe, en anglès *multilingual information retrieval* (MIR)

D'una manera àmplia la **recuperació d'informació** agrupa un conjunt de tècniques que tenen com a objectiu recuperar un objecte en qualsevol mitjà (text, imatge, àudio, vídeo) a partir d'una consulta de l'usuari. En aquest mòdul ens referirem a informació textual i el que pretenen aquestes tècniques és trobar un document o conjunt de documents rellevants a una consulta de l'usuari.

L'**extracció d'informació** és la identificació automàtica d'un tipus seleccionat d'entitats, relacions o esdeveniments en text lliure. Per exemple, aquests sistemes poden intentar recuperar totes les entitats amb nom (noms de persona, d'empreses o institucions, de ciutats, etc.) i classificar-les convenientment.

Els **sistemes de pregunta-resposta** intenten identificar la resposta a una pregunta formulada per l'usuari. Aquests tipus de sistemes no pretenen retornar una sèrie de documents on es troba la resposta a la pregunta formulada, si no el fragment de text que respon a la pregunta.

Els sistemes tradicionals de recuperació d'informació estan dissenyats per a treballar en una única llengua, és a dir, que tant la consulta de l'usuari com la col·lecció de documents amb la qual es treballa estiguin en una única llengua. Les tècniques de **recuperació d'informació multilingüe** pretenen recuperar documents en més d'una llengua a consultes de l'usuari que poden estar escrites en diferents llengües. És a dir, que a partir d'una consulta, per exemple,

en català, recuperi tant documents en català com en castellà, anglès, etc., que siguin rellevants per a la consulta. A més, aquests sistemes intenten presentar la informació en la llengua de l'usuari.

En aquest mòdul dedicarem una secció als cercadors d'Intenet. Aquests sistemes s'enfronten al problema de l'enorme quantitat de documents que indexen (una certa fracció del contingut de la Web), ja que han de retornar els documents rellevants en un temps molt curt.

Objectius

Els objectius bàsics que ha d'haver aconseguit l'estudiant una vegada treballats els continguts d'aquest mòdul són els següents:

1. Comprendre en què consisteix la tasca de recuperació d'informació i conèixer les principals tècniques per a dur-la a terme.

2. Comprendre en què consisteix la tasca d'extracció d'informació, les diferències amb la tasca de recuperació d'informació i conèixer les principals tècniques per a dur-la a terme.

3. Saber diferenciar les finalitats d'uns sistemes de recuperació d'informació i d'un sistema de pregunta-resposta.

4. Conèixer les principals dificultats i les tècniques per a dur a terme les tasques de recuperació i extracció d'informació de manera multilingüe.

5. Conèixer el funcionament dels cercadors d'Internet i els principals serveis que ofereixen.

1. Recuperació d'informació

La **recuperació d'informació** (en anglès, *information retrieval*) consisteix en l'obtenció d'informació continguda en dades emmagatzemades a partir d'una consulta formulada per un usuari.

Aquesta informació pot estar emmagatzemada en qualsevol format o mitjà (text, enregistrament d'àudio, vídeo, bases de dades o en combinació d'aquests), tot i que en aquest apartat tractarem de la recuperació d'informació continguda en documents. Fa un temps l'ús d'eines de recuperació d'informació estava restringit a especialistes (metges, advocats, periodistes, etc.) però en els darrers anys l'ús massiu d'Internet ha popularitzat aquests sistemes. Dedicarem un apartat posterior als cercadors d'Internet, ja que, tot i que en essència aquests sistemes són sistemes de recuperació d'informació, presenten la particularitat que indexen quantitats ingents d'informació i que han de donar respostes en temps molts curts.

Vegeu també

Sobre els cercadors vegeu l'apartat 5 d'aquest mateix mòdul.

1.1. Components d'un sistema de recuperació d'informació

L'objectiu d'un sistema de recuperació d'informació és retornar un conjunt de textos o documents que siguin rellevants a la consulta que ha fet l'usuari. Un primer factor que afectarà molt el disseny d'aquests sistemes serà el tipus de textos o documents que composen el corpus sobre el qual es portaran a terme les cerques. Aquests textos poden ser des de textos molt especialitzats (informes mèdics, lleis o sentències judicials) fins a col·leccions de missatges de correu electrònic. Aquests textos poden presentar o no certa estructuració coneguda *a priori* o estar classificats o no amb certes metadades (descriptors, paraules clau, etc.). Una altra característica important és conèixer *a priori* si tots els documents de la col·lecció estan escrits en una sola llengua o en més d'una.

1.1.1. Preprocessament de documents

Els models clàssics de sistemes de recuperació d'informació funcionen a partir de la indexació dels documents en una sèrie de paraules clau a partir del seu contingut. En principi es considera que els índexs utilitzats donen una indicació del contingut del document. Tradicionalment aquests índexs es confeccionen a partir de les paraules del document, però amb un preprocessat previ per a estalviar espai. Aquest processament consisteix a no fer servir com a índexs paraules funcionals (articles, preposicions, pronoms, etc.), ja que les paraules no funcionals (substantius, adjectius i verbs) són les que contenen la informació semàntica i per tant les que poden representar millor el contingut del document. Quan no es disposa d'informació lingüística per a determinar a

quina categoria gramatical pertany cada paraula s'acostuma a fer la simplificació de considerar que les paraules molt curtes (menys d'un determinat nombre de caràcters, per exemple dos o tres) són paraules funcionals i que les més llargues són paraules amb contingut. A més cal tenir en compte que les paraules funcionals són les més freqüents al document i que eliminar-les ajuda en termes d'eficiència. Tot i això, no hi ha una unanimitat entre els investigadors en aquesta àrea sobre la conveniència o no de l'eliminació de les paraules funcionals per a la representació dels documents (Riloff, 1995).

Una altra tècnica per a reduir el nombre d'índexs és reduir totes les formes flexionades d'una paraula per la seva arrel o per la seva forma base (lema).

Vegeu també

Veurem amb més detall els possibles tractaments per reduir el nombre d'índexs al subapartat 1.2.

Un cop s'han determinat les paraules clau d'un document cal tenir en compte que no totes aquestes paraules seran rellevants per al contingut del document.

> Si una determinada paraula apareix en 1.000 documents de la nostra col·lecció de 5.000 documents, aquesta paraula serà pràcticament inútil per a discriminar un document d'un altre. En canvi, si una altra paraula només apareix en 5 documents, aquesta paraula tindrà molt més poder discriminatori. Si a més, en un d'aquests 5 documents aquesta paraula apareix més vegades, també ponderarà la rellevància d'aquest document.

Així, doncs, podem parlar de dues freqüències diferenciades: la freqüència del terme, és a dir, quantes vegades apareix una paraula en un document o en tota la col·lecció; i la freqüència en documents, és a dir, en quants documents de la col·lecció apareix una determinada paraula.

La **freqüència d'un terme** (en anglès, *term frequency*) és el nombre de vegades que apareix una determinada paraula o grup de paraules en un document o en tota la col·lecció de documents. La **freqüència en documents** (en anglès, *document frequency*) és el nombre de documents on apareix una determinada paraula o grup de paraules.

1.1.2. Construcció de l'índex del document

Per a cada document de la col·lecció s'acostuma a calcular un **índex invertit**, que és una taula on cada paraula o paraula clau del document apareix amb la posició en caràcters en què apareix la paraula. Per exemple, si tenim un document que conté el text següent:

El president del Govern espanyol, José Luis Rodríguez Zapatero, ha participat aquest diumenge en la inauguració de la XIV Cimera de la Unió Africana que se celebra a Addis Abeba, en qualitat de president de torn de la UE. El ministre espanyol d'Afers Estrangers s'ha reunit amb els responsables d'Exteriors de Mauritània, Naha Mint Muknas, i de Mali, Moctar Ouane, per a avaluar les gestions que s'estan realitzant per a intentar alliberar els cooperants catalans. Moratinos no ha volgut comentar el contingut de la reunió i s'ha limitat a afirmar que "continuen treballant", mentre altres fonts de la delegació espa-

nyola reiteraven la necessitat de "mantenir la prudència i la discreció per responsabilitat". Per la seva banda, el ministre de Mali ha afirmat que continuen treballant per A alliberar els segrestats al més aviat possible.

l'índex invertit tindria l'aspecte següent:

a [165, 552]
abeba [173]
addis [167]
afers [250]
afirmar [554]
afirmat [771]
africana [141]
alliberar [441, 808]
altres [599]
amb [281]
aquest [79]
avaluar [382]
aviat [840]
banda [740]
catalans [466]

A continuació presentem un petit programa en Python per a assolir aquest resultat (programa-6-1.py):

Nota

Si no teniu coneixements de programació no és necessari que intenteu entendre com funciona el programa-6-1.py.

```
from nltk.corpus import PlaintextCorpusReader

lector=PlaintextCorpusReader(".","noticia.txt",encoding=" utf-8")
paraules=lector.words()
textcomplet=" ".join(paraules)
index={}

for paraula in paraules:
   if paraula.isalnum():
       posicions=[]
       posicio=0
       while (posicio>-1):
         posicio=textcomplet.find(" "+paraula+" ",posicio+1)
         if (posicio>-1):
             posicions.append(posicio)
       if not(index.has_key(paraula.lower())):
         index[paraula.lower()]=posicions

claus=index.keys()
```

```
claus.sort()
for clau in claus:
    print clau, index[clau]
```

Funcionament del programa-6-1.py

No és necessari que entengueu com funciona exactament el programa. Expliquem breument el seu funcionament per a aquells alumnes amb coneixements de programació. Cal tenir en compte que fem servir la classe *PlaintextCorpusReader* del paquet NLTK (*Natural Language Toolkit*). Creem un objecte amb aquesta classe i el fitxer "noticia.txt", tot indicant que la codificació de l'arxiu és utf-8, a la línia:

```
posicio=textcomplet.find(" "+paraula+" ",posicio+1)
```

Aquest mètode retorna la primera posició on es troba la cadena a cercar (en el nostre cas la paraula amb un espai en blanc al davant i al darrera). Si no troba la cadena retorna un -1. Aprofitem això per posar la cerca en un bucle ja que cada paraula es pot trobar en més d'una posició. Guardem aquestes posicions a una llista i la informació de la paraula passada a minúscula i la llista de posicions a un diccionari. Aquest diccionari és el que realment constitueix l'índex i que imprimim al final del programa. Per a imprimir el que fem és endreçar les claus del diccionari alfabèticament.

Aquesta classe ens proporciona una sèrie de mètodes que farem servir. Fent servir el mètode *words*() obtenim una llista amb totes les paraules que componsen el text (no només les paraules, sinó també els signes de puntuació). Unint els elements d'aquesta llista mitjançant espais en blanc obtenim una cadena de text que representa tot el text, amb l'avantatge que totes les paraules i els signes de puntuació estan separats per un espai en blanc. Al bucle "for paraula in paraules:" recorrem tota la llista de paraules, verifiquem que no siguin signes de puntuació fent servir el mètode *isalnum*() de les cadenes.

Aquest tipus d'índexs es construeixen sobre tots els documents de la col·lecció i els algorismes de cerca i recuperació s'implementen sobre aquests índexs en comptes de fer-ho sobre els mateixos documents.

1.1.3. Processament de les consultes i cerca dels documents que satisfan la consulta

Quan l'usuari d'un sistema de recuperació d'informació fa una consulta, aquesta és processada i es recuperen els documents que són més rellevants. Les tècniques per a identificar els documents més rellevants a partir de la consulta de l'usuari es poden dividir en tres tipus: booleans, models d'espai vectorial i probabilístics.

En els **sistemes booleans** les consultes es representen mitjançant paraules clau connectades per operadors lògics booleans (per exemple, AND, OR o NOT). Aquests tipus de sistemes són molt habituals ja que la semàntica de les representacions booleanes permeten unes implementacions computacionals ràpides i eficients. Un dels inconvenients que presenten aquests sistemes és que recuperen documents basant-se en decisions binàries i no donen a l'usuari una gradació de respostes segons la seva rellevància. Un altre problema important és que, tot i que la semàntica de les consultes booleanes és clara i precisa, els usuaris perden molt de temps transformant les seves necessitats d'informació en expressions booleanes (Belkin i Croft, 1987).

En els **models d'espai vectorial** els documents i les consultes de l'usuari es representen com a vectors (Salton i Lesk, 1968; Salton, 1971). El principal avantatge d'aquests sistemes és que proporcionen un conjunt de documents amb una puntuació associada a cada un d'ells que indica l'índex de rellevància. En el model d'espai vectorial s'assigna uns pesos no binaris a cada índex dels documents i de la mateixa cerca. Aquests pesos es fan servir per a retornar els documents ordenats segons un grau de similitud entre el document i la cerca. A Salton i McGill (1983) podem veure descripcions de diferents tècniques per a calcular els pesos dels índexs. En el model d'espai vectorial es fa servir la mesura anomenada *TF*IDF*. *TF* representa la **freqüència del terme** (en anglès, *term frequency*) i dóna una idea de com un determinat terme descriu el contingut d'un document. *IDF* és la **freqüència inversa en documents** (en anglès, *inverse document frequency*). La idea bàsica és que un terme que apareix en molts documents no és massa útil per a discriminar els documents rellevants dels que no ho són.

En els **models probabilístics** la manera de modelar les consultes d'informació i els documents es basa en la teoria de la probabilitat (Robertson i Spärk Jones, 1976). Atès un conjunt de respostes ideals és possible recuperar el conjunt de documents més propers. Així doncs, es pot pensar que una consulta és una manera d'especificar les propietats del conjunt de respostes ideals. Aquestes propietats es caracteritzen per les propietats semàntiques dels termes d'indexació. No obstant això, les propietats de la resposta ideal no sempre són conegudes en el moment de fer la consulta i és necessari aproximar la resposta al conjunt de propietats més properes. El principal avantatge del model probabilístic és que els documents que retorna el sistema estan ordenats segons la seva probabilitat de ser rellevants. Els inconvenients són que el sistema ha d'endevinar una separació inicial entre els documents rellevants i no rellevants i que aquesta aproximació no té en compte la freqüència dels termes d'indexació dins d'un document.

1.1.4. Avaluació de sistemes de recuperació d'informació

Hi ha un seguit de mesures que es fan servir per a avaluar sistemes de recuperació d'informació. Per a poder aplicar aquestes mesures cal disposar d'un conjunt de documents i d'una consulta i a més saber *a priori* quins documents són rellevants o no a la consulta.

Des de fa uns anys s'organitzen una sèrie de conferències i competicions entre sistemes que proporcionen un conjunt de dades que permeten avaluar els sistemes de recuperació d'informació. Una d'aquestes conferències és *Text Retrieval Conference* (TREC).

Precision

La **precisió** (*precision*) és la relació entre els documents recuperats que són rellevants respecte el total de document recuperats.

$$\text{precision} = \frac{\left|\{\text{documents rellevants}\} \cap \{\text{documents recuperats}\}\right|}{\{\text{documents recuperats}\}}$$

Recall

El *recall* és la fracció de documents que són rellevants i s'han recuperat.

$$\text{recall} = \frac{\left|\{\text{documents rellevants}\} \cap \{\text{documents recuperats}\}\right|}{\{\text{documents rellevants}\}}$$

És molt fàcil obtenir un *recall* del 100% (l'ideal) simplement retornant tots els documents de la col·lecció (així ens assegurem que retornem també tots els rellevants). Així, doncs, aquesta mesura per si mateixa no dóna gaire informació i cal combinar-la amb alguna altra que ens mesuri d'alguna manera el nombre de documents no rellevants (per exemple, combinar *precision* i *recall*).

Fall-Out

És la proporció de documents no rellevants que es recuperen respecte del total de documents no rellevants.

$$\text{fall-out} = \frac{\left|\{\text{documents no rellevants}\} \cap \{\text{documents recuperats}\}\right|}{\{\text{documents no rellevants}\}}$$

També és molt fàcil assolir un *fall-out* del 0% (l'ideal) simplement no retornant cap document.

F-measure

Com hem vist a les mesures anteriors, valors molts alts d'una d'elles pot no significar gran cosa. Per exemple, podem tenir una precisió molt alta retornant molt pocs documents. Pensem en una col·lecció on hi ha 100 documents rellevants a la nostra consulta, podem retornar només un amb molta seguretat i obtenir una precisió del 100% (l'ideal), però amb un *recall* molt baix (1%). La *F-measure* combina precisió i cobertura en una única mesura: és la mitjana harmònica ponderada de precisió i cobertura i es calcula:

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

1.2. Tècniques de processament del llenguatge natural aplicades a la recuperació d'informació

Hi ha hagut un interès important a incorporar tècniques de processament del llenguatge natural dins dels sistemes de recuperació d'informació. La clau d'aquesta incorporació és que les tècniques que s'incorporin han de ser molt robustes i eficients per a poder tractar grans quantitats d'informació sense alentir massa el sistema. Tampoc queda clar en quin nivell o nivells dels sistemes de recuperació d'informació cal incorporar les tècniques de processament del llenguatge natural: en la indexació dels documents, en el processament de les consultes de l'usuari o en la cerca dels documents que satisfan aquestes consultes.

Es pot trobar una bona introducció de les tècniques de processament del llenguatge natural a la indexació de documents per a la recuperació d'informació a Spärk Jones (1999). En aquest article es parla d'indexació motivada lingüísticament (en anglès, *linguistically motivated indexing* o *LMI*).

Vegeu també

En els subapartats següents veurem com tres diferents nivells d'anàlisi (morfològic, sintàctic i semàntic) poden ajudar a la indexació de documents.

1.2.1. Millora de la indexació fent servir informació morfològica

En el programa-6-1.py del subapartat 1.1.2. hem construït uns índexs d'un document basant-nos en les formes de les paraules, és a dir, indexaven com a diferents les paraules

```
sospitosos [2115, 3280]
sospitós [3318]
```

Una primera idea per a millorar la indexació seria unir totes les formes d'una mateixa paraula en una forma base. Per a aconseguir-ho disposem de tres mecanismes:

1) ***Stemming***: és un mecanisme per reduir una paraula a la seva arrel o tema (*stem*). El tema és la part que resta d'un mot quan se'n separen els morfemes flexionals. Per exemple, de *sospitós* i *sospitosos* el tema seria *sospit*.

2) Anàlisi morfològica: aquesta anàlisi torna la forma base o lema i una etiqueta morfosintàctica. De l'exemple *sospitós*, *sospitosos* retornaria *sospitós* i una etiqueta del tipus "AQ0MP0". Quan l'anàlisi retorna totes les possibles lectures d'una determinada paraula parlem pròpiament d'anàlisi morfològica. Quan el sistema és capaç de determinar quina etiqueta i lema són els correctes per a paraules amb més d'una lectura, parlem d'**anàlisi morfosintàctica**.

3) Lematització: és similar a l'anàlisi morfològica però en aquest cas només retorna el lema associat.

Un dels primers algorismes eficients per a portar a terme la tasca d'*stemming* és el de Porter (1980). Al programa-6-2.py llegim una notícia en anglès i apliquem l'*stemmer* de Porter que es distribueix amb l'NLTK.

```
from nltk.corpus import PlaintextCorpusReader
from nltk.stem.porter import PorterStemmer

lector=PlaintextCorpusReader(".","noticia-eng.txt")
paraules=lector.words()

for paraula in paraules:
    print paraula, PorterStemmer().stem(paraula)
```

I a continuació podem veure un fragment de la sortida:

```
The The
imbalances imbal
built built
up up
during dure
the the
previous previou
lengthy lengthi
, ,
robust robust
upturn upturn
will will
continue continu
to to
weigh weigh
down down
on on
activity activ
in in
2010 2010
and and
```

```
2011 2011
```

L'anàlisi morfològica retorna el lema de la paraula i una etiqueta morfosintàctica. Si una paraula té més d'una possible lectura les retorna totes. A continuació observem la sortida de l'analitzador Freeling per a la mateixa frase de la notícia:

Web recomanat

http://garraf.epsevg.upc.es/freeling/demo.php

```
The the DT 1
imbalances imbalance NNS 1
built build VBN 0.831818 build VBD 0.168182
up up RP 0.523547 up RB 0.310904 up IN 0.16531 up NN 7.98212e-05 up VB 7.98212e-05 up
VBP 7.98212e-05
during during IN 1
the the DT 1
previous previous JJ 1
lengthy lengthy JJ 1
, , Fc 1
robust robust JJ 1
upturn upturn NN 0.762874 upturn VB 0.181391 upturn VBP 0.0557355
will will MD 0.991197 will NN 0.00843388 will VB 0.000307806 will VBP 6.15612e-05
continue continue VB 0.812893 continue VBP 0.187107
to to TO 0.999909 to IN 9.13109e-05
weigh weigh VB 0.576923 weigh VBP 0.423077
down down RB 0.506688 down RP 0.278923 down IN 0.210593 down NN 0.00234996 down VBP 0.00126537
down...
on on IN 0.972791 on RP 0.0217369 on RB 0.00547189
activity activity NN 1
in_2010 [??:??/??/2010:??.??:??] W 1
and and CC 1
2011 2011 Z 1
. . Fp 1
```

Fixem-nos que Freeling ens retorna cada una de les possibles lectures de cada paraula i la seva probabilitat. En els casos que només retorna una lectura la probabilitat és 1.

En canvi, l'anàlisi morfosintàctica retorna únicament una lectura per a cada paraula:

```
The the DT
imbalances imbalance NNS
built build VBN
up up RP
during during IN
the the DT
previous previous JJ
lengthy lengthy JJ
```

```
, , Fc
robust robust JJ
upturn upturn NN
will will MD
continue continue VB
to to TO
weigh weigh VB
down down RP
on on IN
activity activityNN
in_2010 [??:??/??/2010:??.??:??] W
and and CC
2011 2011 Z
. . Fp
```

Finalment, al programa-6-3.py mostrem el funcionament d'un lematitzador que es distribueix amb el paquet NLTK:

```
from nltk.corpus import PlaintextCorpusReader
from nltk.stem.wordnet import WordNetLemmatizer

lector=PlaintextCorpusReader(".","noticia-eng.txt")
paraules=lector.words()

for paraula in paraules:
    print paraula, WordNetLemmatizer().lemmatize(paraula)
```

i la seva sortida:

```
The The
imbalances imbalance
built built
up up
during during
the the
previous previous
lengthy lengthy
, ,
robust robust
upturn upturn
will will
continue continue
to to
weigh weigh
down down
on on
activity activity
```

```
in in
2010 2010
and and
2011 2011
. .
```

1.2.2. Millora de la indexació fent servir informació sintàctica

Hi ha hagut diversos intents d'incorporar informació sintàctica en un sistema de recuperació d'informació, però en general les millores assolides han estat poc significatives. En Dillon i Gray (1983) s'ha investigat l'ús de patrons sintàctics, a Bruandet (1985) i Kerkouba (1985) es fan servir les categories sintàctiques per a determinar dependències i a Bartschi (1984) es determinen frases en el context d'una llista coneguda d'*stop-words* i restriccions adjacents. Cap d'aquestes propostes han tingut gaire èxit i totes elles tenen en comú que fan servir la informació sintàctica per a la indexació de documents.

En canvi, a Croft (1985) i Smeaton (1986) s'inclou la informació sintàctica en el moment de la recuperació i en termes d'una consulta particular. També a Mittendorfer i Winiworter (2002) es presenta un sistema que porta a terme l'anàlisi sintàctica de les consultes dels usuaris per a millorar un sistema de recuperació d'informació. Dels experiments que presenta es dedueix que aquest mètode millora els resultats només en uns tipus de consultes molt concretes.

1.2.3. Millora de la indexació fent servir informació semàntica

La idea bàsica és millorar la qualitat dels sistemes afegint informació sobre el significat de les paraules i fent servir aquest coneixement per desambiguar les paraules i identificar relacions entre les paraules. S'han proposat diverses tècniques per a aprofitar la informació semàntica en el procés de recuperació d'informació. Cal fer una distinció entre els sistemes de recuperació desenvolupats per a un domini específic dels sistemes desenvolupats per a tractar amb text lliure (Kraaij i Pohlmann, 1996).

Els sistemes de recuperació d'informació desenvolupats per a un domini específic (per exemple, medicina, dret, etc.) es basen en la suposició que determinats significats d'una paraula polisèmica no hi apareixeran en el domini o hi apareixeran molt poc. En aquest tipus de sistemes es fan servir tesaurus construïts de manera manual o automàtica que representen els conceptes del domini i les relacions entre aquests conceptes (sinonímia, antonímia, hiponímia, etc.). Aquests tesaurus es fan servir tant en la indexació dels documents com en les expansions de les consultes dels usuaris (per exemple, afegint termes relacionats a la consulta). Podeu trobar exemples d'aquests tipus de sistemes a Salton (1989). Aquest tipus de desambiguació que elimina alguns sentits de les paraules sobre la base del domini sovint rep el nom de *desambiguació global*.

En canvi els sistemes que tracten amb text lliure, és a dir, no restringit a un determinat domini, han de basar-se en tècniques de *desambiguació local*, és a dir, que es basen en el context d'aparició immediat. Aquestes tècniques reben el nom genèric de *Word Sense Desambiguation*. En general, es poden distingir quatre tipus de sistemes de desambiguació del sentit de les paraules:

1) Mètodes basats en diccionaris o en altres tipus de coneixement. Es basen principalment en l'ús de diccionaris, tesaurus i bases de coneixement lèxic, sense fer servir corpus. L'algorisme de Lesk (Lesk, 1986) és l'exemple prototípic d'aquest tipus d'algorismes. Es basa en la hipòtesi que les paraules que es fan servir en un mateix text estan relacionades entre elles i que aquesta relació també es pot observar en les definicions d'un diccionari. Dues paraules o més es poden desambiguar buscant les seves definicions a un diccionari i observant quina definició té més paraules que coocorren en el text. Una alternativa a l'ús de definicions de diccionaris és fer servir bases de dades de coneixement lèxic com ara Wordnet (Miller, 1995) i calcular la similitud semàntica de cada parell de sentits de les paraules.

2) Mètodes aprenentatge automàtic supervisat. Fan ús de corpus anotats semànticament per a entrenar sistemes d'aprenentatge automàtic. La majoria d'algorismes d'aprenentatge automàtic s'han fet servir per a la tasca de *Word Sense Disambiguation* (Escudero i altres, 2000).

3) Mètodes semisupervisats. Aquest tipus d'algorismes poden fer servir tant dades anotades com no anotades i han sorgit per la dificultat d'obtenir corpus anotats semànticament. Un dels primers algorismes d'aquest tipus va ser l'algorisme de Yarowsky (Yarowsky, 1995). Aquest algorisme es basa en l'observació que la majoria de paraules tendeixen a tenir un únic sentit en un determinat discurs i col·locació. En aquest grup també es poden agrupar molts altres algorismes que fan servir la tècnica de *bootstrapping*. Aquesta tècnica consisteix a fer servir un corpus anotat petit per a entrenar un sistema automàtic. El sistema entrenat es fa servir per a anotar un corpus més gran. D'aquest corpus ens quedem només amb els fragments que han estat anotats de manera més fiable. Aquesta fracció més segura del corpus anotat automàticament es fa servir per a entrenar de nou un sistema automàtic. Aquest procés es repeteix fins que es considera que s'assoleixen uns nivells de precisió acceptables. En Wang i Hoffmann (2006) es presenta un mètode que fa servir dades dinàmiques de la Web obtingudes mitjançant motors de cerca d'Internet per a enriquir mitjançant *bootstrapping* el coneixement semàntic d'un sistema de *Word Sense Disambiguation*.

4) Mètodes sense supervisar. Són mètodes que fan servir de manera gairebé exclusiva corpus sense anotar. La hipòtesi bàsica de treball és que els sentits similars ocorren en contextos similars i que els sentits es poden induir a partir de textos fent servir tècniques de clusterització de paraules fent servir algun tipus de mesura de similitud de context (Papp, 2009).

1.3. GATE com a sistema de recuperació d'informació

La *General Architecture for Text Engineering* (GATE) (Cunningham, 2000) és una arquitectura de programari per a l'enginyeria lingüística. És a dir, és un entorn de desenvolupament per a aplicacions de processament del llenguatge natural.

Vegeu també

En el subapartat 2.4. veurem les funcionalitats d'extracció d'informació del sistema GATE.

En aquest subapartat farem servir GATE per a fer una demostració d'un sistema de recuperació d'informació en funcionament. Veurem GATE funcionant amb aquestes tasques, però cal tenir en compte que GATE és molt més que això, ja que permet desenvolupar les nostres pròpies aplicacions. Perquè us feu una idea, el manual d'usuari té més de 500 pàgines.

Web recomanat

Tot el que fa referència a GATE (instal·lables dels programes, manuals, demostracions, vídeos explicatius, etc.) es pot obtenir a la pàgina web http://gate.ac.uk. Us animem a descarregar la versió corresponent al vostre sistema operatiu per a poder fer vosaltres mateixos les demostracions que oferim a continuació.

GATE és una aplicació de programari lliure amb llicència GPL i pot funcionar sota Windows, Linux i Mac.

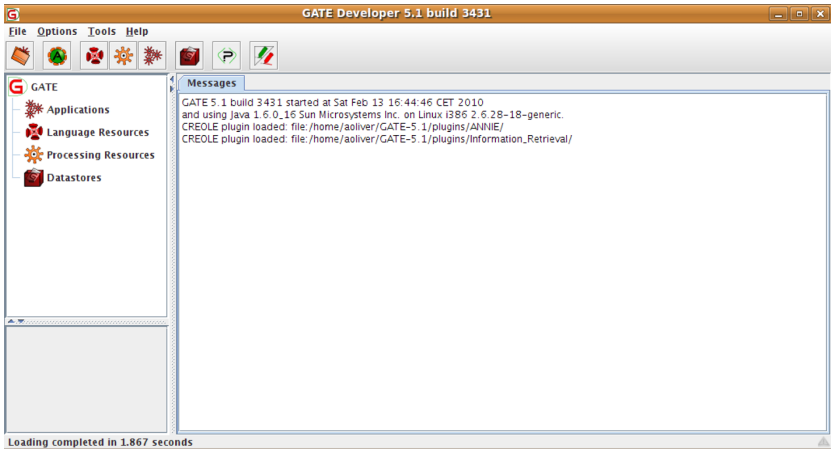
1.3.1. Instal·lació i execució de GATE

GATE està escrit en Java, així que abans d'instal·lar-lo us heu d'assegurar que tingueu el Java correctament instal·lat al vostre ordinador. Un cop verificat això descarregueu el programa d'instal·lació corresponent al vostre sistema operatiu:

- Per a Linux l'arxiu és un JAR de Java (que podeu executar fent `java -jar`).
- Per a Windows és un arxiu .exe que podeu executar fent doble clic damunt seu un cop descarregat.
- Per a Mac és un arxiu .dmg que s'instal·la de la manera habitual.

Un cop instal·lat heu d'executar l'aplicació "GATE GUI" que s'haurà ubicat en algun menú del sistema, depenent del sistema operatiu. Apareixerà una pantalla com la de la figura 1.

Figura 1. Pantalla inicial del programa GATE-GUI



En aquest subapartat indexarem un petit corpus de 10 articles de la Viquipèdia. Primer farem la prova per a l'anglès i després per al català. Tot el procés consta dels passos següents:

1) Activar el connector de recuperació d'informació.

2) Crear el corpus.

3) Crear un *serial datastore* i assignar-li el corpus.

4) Indexar el corpus.

5) Crear un recurs de processament per a la cerca.

6) Crear una *pipeline application* que contingui el recurs de processament.

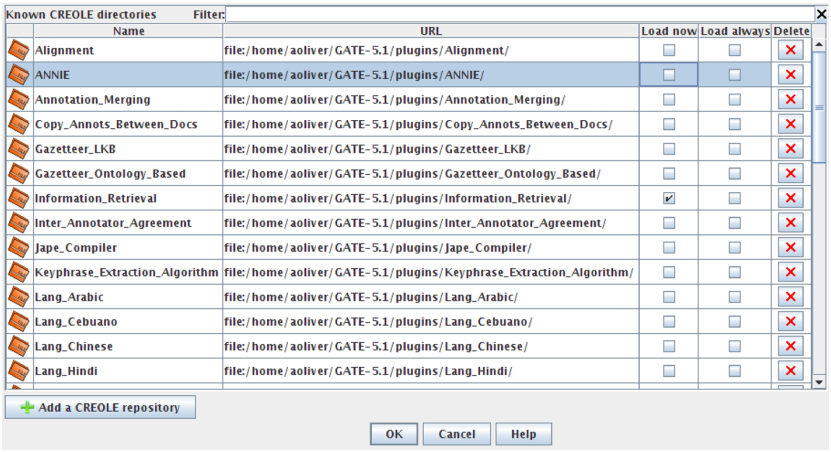
7) Fer cerques.

En els subapartats següents expliquem amb detall cada un d'aquests passos.

1.3.2. Activació del connector de recuperació d'informació

GATE disposa d'un gran nombre de connectors per a la realització de tasques de processament del llenguatge natural. Un d'aquests connectors permet fer tasques relacionades amb la recuperació d'informació. Per activar aquest connector heu de fer *File* > *Manage CREOLE Plugins*. Apareixerà una pantalla com la de la figura 2. Seleccioneu la casella *Load Now* corresponent al connector *Information_Retrieval* i feu clic al botó OK.

Figura 2. Pantalla de gestió dels connectors CREOLES



1.3.3. Creació del corpus

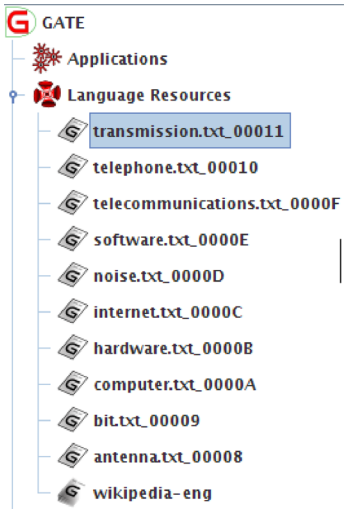
Hi ha diferents maneres de crear un corpus amb GATE. La que explicarem és la més adequada si volem crear un corpus i assignar-li un conjunt de documents que hi ha en un directori. En la finestra de l'esquerra seleccioneu *Language*

Resources i feu clic al botó dret del ratolí i seleccioneu *New GATE Corpus*, a la pantalla que hi apareix (figura 3) n'hi ha prou a especificar un nom per al corpus.

Figura 3. Pantalla de creació d'un nou corpus

Un cop creat afegirem el directori "wikipedia-eng", que conté 10 articles de la Viquipèdia en anglès situant-nos sobre el corpus que acabem de crear a la pantalla de l'esquerra i fent clic al botó dret del ratolí. Seleccionem l'opció *Populate* i en la pantalla que hi apareix seleccionem el directori que conté els arxius fent clic al botó OK. Un cop fet aquest pas en la pantalla de l'esquerra apareixeran els noms de tots els fitxers que hem carregat i del mateix corpus, tal com podem observar a la figura 4.

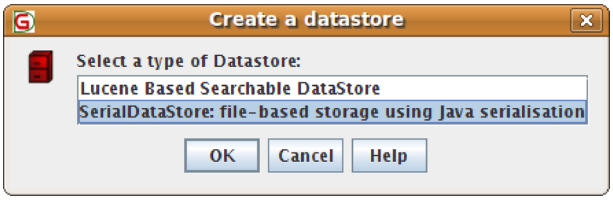
Figura 4. Recursos lingüístics carregats al sistema



1.3.4. Creació del *datastore* i assignació del corpus al *datastore*

Únicament és possible indexar els corpus que estan dins d'un *datastore* del tipus *SerialDataStore*. Per aquest motiu serà imprescindible crear-lo i assignar-li el corpus. Per a fer això ens situem sobre *Datastores* a la pantalla de l'esquerra i fem clic al botó dret del ratolí i seleccionem *New*. Apareixerà una pantalla com la de la figura 5 on haurem de seleccionar *SerialDataStore* i fer clic a OK. A continuació apareixerà un diàleg que ens demanarà que creem un directori nou. Un cop fet això, a la pantalla de l'esquerra apareixerà el nou *Datastore* que hem creat.

Figura 5. Selecció del tipus de *datastore*



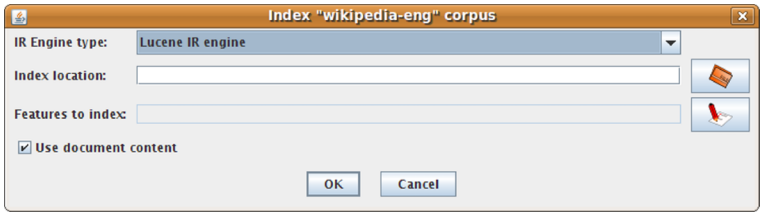
Ara hem d'assignar el corpus que hem creat al pas anterior a aquest *Datastore*. Per a fer això ens situem sobre el corpus a la pantalla de l'esquerra i fem clic al botó dret del ratolí. Apareixerà un menú on seleccionarem l'opció *Save to datastore* i apareixerà una pantalla com la de la figura 6 on seleccionarem el *datastore* que acabem de crear i farem clic al botó OK.

Figura 6. Selecció del *datastore*

1.3.5. Indexació del corpus

Ara ja podem indexar el corpus des del *datastore*. Per a fer això simplement ens hem de situar sobre el corpus en la pantalla de l'esquerra, fer clic amb el botó dret del ratolí i seleccionar l'opció *Index Corpus*. Apareixerà una pantalla com la que es mostra a la figura 7. En aquesta pantalla indicarem on s'ubicaran els índexs amb *Index Location*. Un cop indicada aquesta ubicació començarà la indexació del corpus. Aquest procés, per a corpus grans, pot trigar una mica. Un cop acabada la indexació surt un missatge a la barra d'estat que indica que el corpus ja està indexat i el temps que ha trigat a fer-ho.

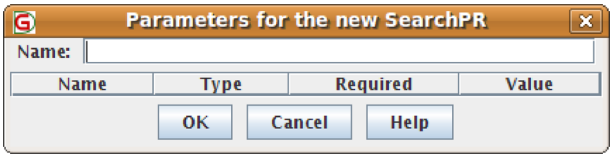
Figura 7. Selecció de les opcions d'indexació



1.3.6. Creació d'un recurs de processament per a la cerca

Ara ja tenim el corpus indexat i hem de crear el recurs de processament per a fer les cerques. Per a fer això ens situem sobre *Processing Resources* a la pantalla de l'esquerra, fem clic amb el botó dret del ratolí i seleccionen *New* i *SearchPR*. Apareixerà una pantalla com la de la figura 8 on podem indicar les propietats d'aquest recurs de processament. Ara ens limitarem a donar-li un nom.

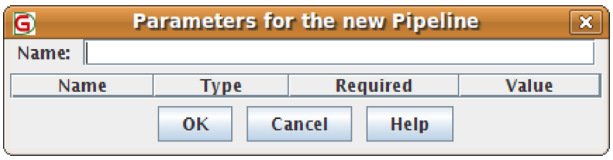
Figura 8. Selecció de les propietats del *SearchPR*



1.3.7. Creació del *pipeline* application

Ara hem de seleccionar *Applications* a la pantalla de l'esquerra i fer clic en el botó dret del ratolí i seleccionar *Pipeline*. A la pantalla que hi apareix (figura 9) indicarem un nom.

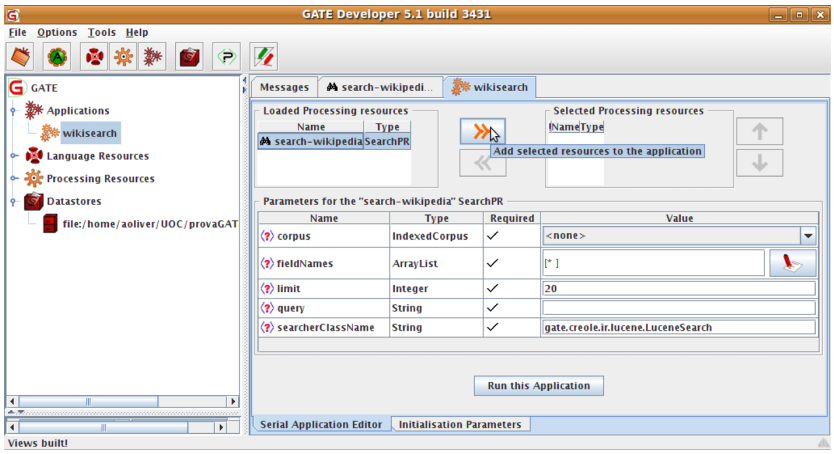
Figura 9. Creació del *Pipeline*



1.3.8. Realització de la cerca

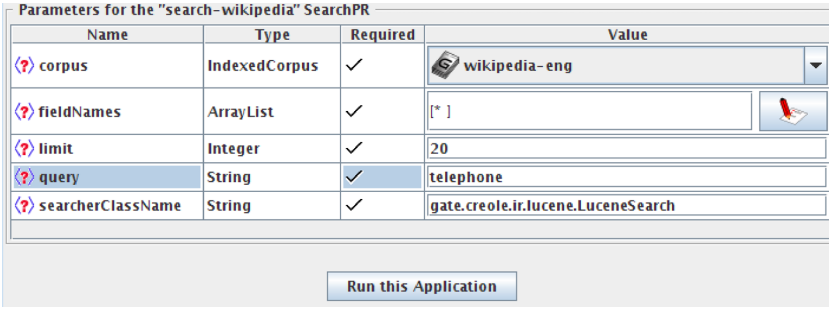
Un cop fet això farem doble clic en el *Pipeline* creat i apareixerà una pantalla com la de la figura 10 on haurem de seleccionar el recurs de procés creat i passar-lo a la part de la dreta amb la fletxa.

Figura 10. Pantalla de selecció del recurs de procés



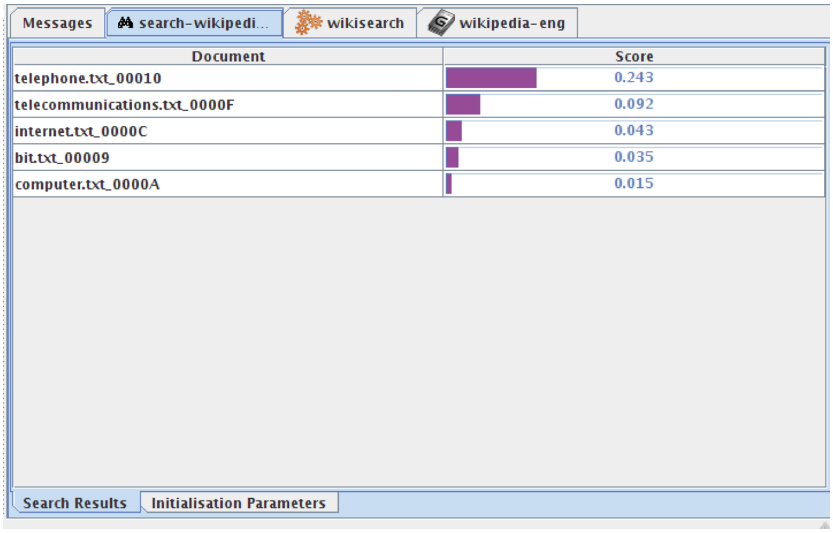
Abans de passar-lo a la part dreta haurem de seleccionar els paràmetres, sobretot el corpus i la pròpia consulta (figura 11).

Figura 11. Selecció dels paràmetres de la cerca



Un cop fet això podem passar-lo a la dreta i fer clic al botó *Run Application* i apareixeran els resultats si fem clic en el recurs de procés *SearchPR* i seleccionem la pestanya *Search Results* (figura 12).

Figura 12. Resultats de la cerca



1.4. Conclusions

En aquest apartat hem presentat els conceptes bàsics de la tasca anomenada recuperació d'informació (*information retrieval*). Aquesta tasca pretén recuperar una sèrie de documents rellevants a partir d'una consulta formulada per l'usuari. El sistema es limita a presentar un conjunt de documents, normalment endreçats per rellevància, que responen a la necessitat d'informació de l'usuari. L'usuari haurà d'explorar aquests documents i trobar les seccions rellevants per a la seva consulta.

El conjunt de documents sobre el qual es faran les consultes s'han d'indexar prèviament, de manera que la cerca sobre aquests documents es faci d'una manera efectiva. El sistema també ha de processar la consulta de l'usuari i disposar d'algun algorisme que permeti cercar els documents rellevants d'una manera ràpida. Hem vist que aquest tipus de sistemes poden funcionar bé sense fer ús de tècniques específiques de processament del llenguatge natural.

Hem vist també algunes tècniques de processament del llenguatge natural que poden ajudar en certa mesura a la recuperació d'informació. Aquestes tècniques es poden aplicar tant durant el procés d'indexació, com en el processament de la petició de l'usuari. Algunes tècniques fins i tot s'apliquen únicament sobre els documents retornats pel sistema de recuperació. Les tècniques de processament del llenguatge natural poden funcionar en diferents nivells d'anàlisi: morfològic, sintàctic o semàntic.

2. Extracció d'informació

L'extracció d'informació és un procés que pren com a entrada un conjunt de textos i produeix la sortida dades no ambigües en un format donat. Aquestes dades es poden mostrar directament als usuaris o bé poden ser emmagatzemades en bases de dades o fulls de càlcul per a portar a terme anàlisis posteriors.

L'**extracció d'informació** (en anglès, *information extraction*) és la identificació automàtica de tipus seleccionats d'entitats, relacions o esdeveniments en text lliure. Cobreix un gran nombre de tasques, des de trobar tots els noms d'empreses en un text, fins a trobar tots els assassins i qui va matar a qui, quan i on (Grishman, 2003).

Un sistema de recuperació d'informació com els que hem vist a l'apartat 1 troba un conjunt de textos que satisfan la consulta de l'usuari. En canvi, els sistemes d'extracció d'informació analitzen textos i presenten únicament la informació que és rellevant per a la consulta de l'usuari.

Si estem interessats en l'evolució del preu de les accions d'una determinada empresa, un sistema de recuperació d'informació detectaria els documents on es parla dels preus de les accions d'aquesta empresa i els mostraria a l'usuari. Llavors l'usuari hauria de llegir els documents i recopilar la informació sobre els preus. En canvi, un sistema d'extracció d'informació retornaria els preus de les accions en alguna estructura de dades que permetés una anàlisi posterior.

En aquest apartat veurem diferents tasques relacionades amb l'extracció d'informació:

- **Reconeixement d'entitats amb nom** (en anglès, *Named Entity Recognition,* NER): és el reconeixement d'entitats amb nom, com el nom de persones, empreses i institucions, noms de lloc, expressions temporals i alguns tipus d'expressions numèriques. A més del reconeixement aquesta tasca inclou també la classificació d'aquestes entitats.

- **Resolució de la coreferència** (en anglès, *Coreference Resolution*, CO): identifica relacions d'identitat entre entitats que apareixen als textos.

Exemples

Un exemple és la relació entre una entitat amb nom i un determinat pronom. Per exemple, si en el text apareix "Felipe González diu que tant ell com Aznar i Zapatero van ordenar parlar amb ETA" el sistema hauria de poder relacionar "ell" i "Felipe González". Un altre tipus de relació que han de solucionar aquests sistemes és el de les entitats amb nom amb les seves sigles. Per exemple, si tenim les oracions "Convergència i Unió ha creat una pàgina web de la Casa Gran del Catalanisme a Lleida, estructurada com una casa on a les diferents estances de la casa s'hi pot trobar espais d'opinió de tothom que s'hi adhereixi i poder fer difusió de les activitats i les idees que aporti la gent. Paral·lelament al web, CiU ha creat un grup al Facebook que ja té 176 membres." el sistema hauria de relacionar "Convergència i Unió" amb "CiU".

- **Extracció d'esdeveniments** (en anglès, *Event Extraction*): pretén detectar certs tipus d'esdeveniments i la seva informació associada, com per exemple compres d'empreses amb la informació de qui compra a qui, quan, per quants diners, etc.

D'un text com el següent: "El 25 d'octubre es va formalitzar l'acord, i Telco -un grup liderat per Telefónica en què també hi ha diferents bancs- va comprar la societat Olimpia, un hòlding de Pirelli i l'empresa Sintonia, propietat de la família Benetton, que tenia el 23,59% de les accions de Telecom Italia. Des de llavors, Telefónica n'és l'accionista majoritari i ha impulsat el canvi de dirigents." El sistema hauria de deduir que el grup Telco va comprar Olimpia. Podria deduir també que la majoria d'accions de Telco són de Telefònica, etc.

Mentre que el reconeixement d'entitats amb nom i l'extracció d'esdeveniments poden tenir un interès per als usuaris, la resolució de la coreferència es considera una tasca auxiliar per a poder millorar els resultats d'altres tasques d'extracció d'informació.

Nota

Es poden trobar moltes més tasques associades a l'extracció d'informació, però en aquest apartat només tractarem les exposades anteriorment.

2.1. Reconeixement d'entitats amb nom

El **reconeixement d'entitats amb nom** (en anglès, *Named Entity Recogntion*, NER) és una subtasca de l'extracció d'informació que intenta localitzar una sèrie d'elements d'un text i classificar-los en una sèrie de categories predefinides com noms de persones, d'empreses o institucions, quantitats, valors monetaris, percentatges, etc.

El reconeixement d'entitats amb nom també rep el nom d'**identificació d'entitats** (*entity identification*) o **extracció d'entitats** (*etity extraction*).

Un sistema de reconeixement d'entitats amb nom si té com a entrada el text següent:

```
Spain said Monday it was willing to take in five inmates from the American prison
in Guantanamo Bay, Cuba, three more than it had announced last month. Foreign
Minister Miguel Angel Moratinos said their nationalities would be announced when
they arrived. Spain had previously agreed to accept a Yemeni and a Palestinian.
```

Retornaria una sortida com la següent:

```
[LOC Spain ] said Monday it was willing to take in five inmates from the [MISC
American ] prison in [LOC Guantanamo Bay ] , [LOC Cuba ] , three more than it
had announced last month . Foreign Minister [PER Miguel Angel Moratinos ] said
their nationalities would be announced when they arrived . [LOC Spain ] had
previously agreed to accept a [MISC Yemeni ] and a [MISC Palestinian ].
```

És a dir, ens ha detectat com a noms de lloc (LOC): Spain, Guantanamo Bay, Cuba; com a noms de persona: Miguel Angel Moratinos i com a altres (MISC): America i Palestinian.

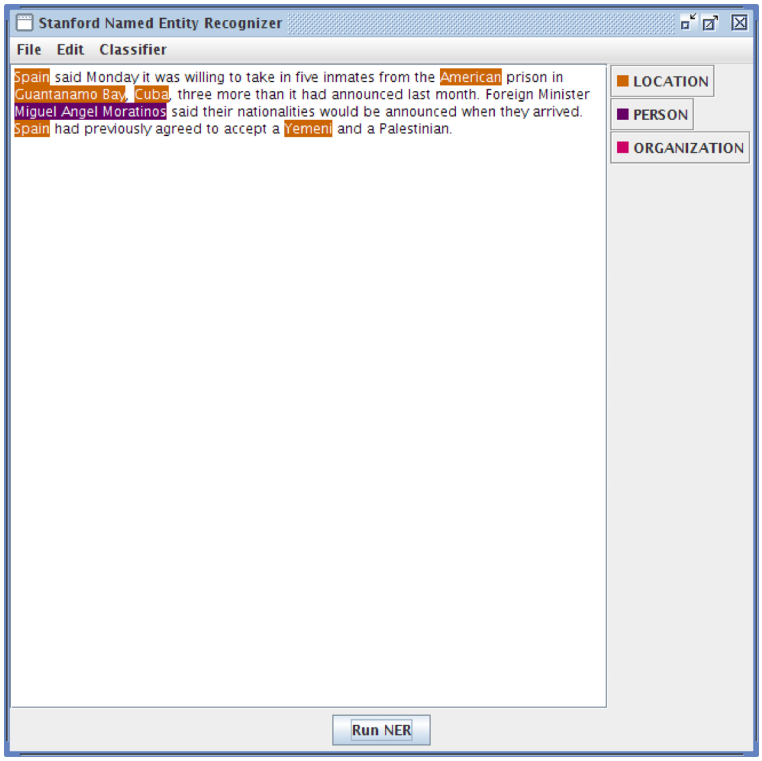
El sistema que hem fet servir és l'LBJ Named Entity Tagger (Ratinov i Roth, 2009). Un altre sistema que, com aquest, està disponible per a l'anglès i es pot descarregar és l'Stanford Named Entity Recogniser (NER) (Frinkel i altres, 2005). El sistema es pot descarregar des d'Internet. Un cop descarregat i comprimit es pot executar el programa amb "./ner-gui.sh" des de Linux o "ner.bat" des de Windows. Un cop s'obre l'aplicació hem de fer *Classifier* > *Load Default CRF* (per a carregar el sistema per defecte). Un cop fet això ja podem carregar un text per a analitzar amb el menú *File* o simplement enganxant el text que volem analitzar. A la figura 13 podem observar l'anàlisi del mateix text amb aquesta eina.

Webs recomanats

Es pot trobar una demo de l'LBJ Named Entity Tagger al web http://l2r.cs.uiuc.edu/cogcomp/LbjNer.php i també es pot descarregar l'aplicació. Malauradament només funciona per l'anglès.

L'Stanford Named Entity Recogniser (NER) es pot descarregar des de http://nlp.stanford.edu/software/CRF-NER.shtml.

Figura 13. Stanford Named Entity Recogniser



Si fem *File* > *Save Tagged File As* podem guardar el resultat en un arxiu que tindrà l'aspecte següent:

```
<LOCATION>Spain</LOCATION> said Monday it was willing to take in five inmates
from the <LOCATION>American</LOCATION> prison in <LOCATION>Guantanamo
Bay</LOCATION>, <LOCATION>Cuba</LOCATION>, three more than it had
announced last month. Foreign Minister <PERSON>Miguel Angel
Moratinos</PERSON> said their nationalities would be announced when they
arrived. <LOCATION>Spain</LOCATION> had previously agreed to accept a
<LOCATION>Yemeni</LOCATION> and a Palestinian.
```

El sistema que ve per defecte està dissenyat per a l'anglès, però pot funcionar més o menys bé per a altres llengües. Per exemple, si analitzem amb aquest programa el fragment següent d'una notícia en català:

Mas creu que Maragall continua pensant que el tripartit no té projecte de país tot i la rectificació considera que les reflexions del conseller d'Educació estaven "molt pensades".

ACN

Barcelona

Ult. Act. 16/02/2010 13:17

El president de CiU, Artur Mas, considera que si el conseller d'Educació, Ernest Maragall, va escriure en un article que el tripartit no té projecte de país, és perquè realment ho pensa, tot i la rectificació.

En declaracions a Onda Rambla-Punto Radio, ha assenyalat que "és molt significatiu" que Maragall ressaltés la falta de projecte de país, i ha dit que era una reflexió molt pensada perquè així ho va deixar per escrit. Mas ha fet aquestes declaracions un dia després que Maragall posés el seu càrrec a disposició del president de la Generalitat, José Montilla, després de les seves polèmiques reflexions. Montilla no li va acceptar la dimissió i Maragall va haver de rectificar.

"Sobre què passa dins del PSC no comentaré res, però és una veritat com un temple i un secret de domini públic el que va dir Maragall, i és que el tripartit no té projecte de país", ha remarcat. Mas ha dit que el conseller no ho va dir, sinó que "ho va escriure i més d'una vegada i estava molt pensat".

Obtenim el resultat següent:

```
Mas creu que <PERSON>Maragall</PERSON> continua pensant que el tripartit no
té projecte de país tot i la rectificació. Considera que les reflexions del conseller
d'Educació estaven "molt pensades".

El president de CiU, <PERSON>Artur Mas</PERSON>, considera que si el conseller
d'Educació, <PERSON>Ernest Maragall</PERSON>, va escriure en un article que el
tripartit no té projecte de país, és perquè realment ho pensa, tot i la rectificació.
```

```
En declaracions a <ORGANIZATION>Onda Rambla-Punto
Radio</ORGANIZATION>, ha assenyalat que "és molt significatiu" que
<ORGANIZATION>Maragall</ORGANIZATION> ressaltés la falta de projecte de
país, i ha dit que era una reflexió molt pensada perquè així ho va deixar per escrit.
Mas ha fet aquestes declaracions un dia després que
<PERSON>Maragall</PERSON> posés el seu càrrec a disposició del president de la
<ORGANIZATION>Generalitat</ORGANIZATION>, <PERSON>José
Montilla</PERSON>, després de les seves polèmiques reflexions.
<LOCATION>Montilla</LOCATION> no li va acceptar la dimissió i
<PERSON>Maragall</PERSON> va haver de rectificar.

"Sobre què passa dins del <ORGANIZATION>PSC</ORGANIZATION> no
comentaré res, però és una veritat com un temple i un secret de domini públic el
que va dir <PERSON>Maragall</PERSON>, i és que el tripartit no té projecte de
país", ha remarcat. Mas ha dit que el conseller no ho va
dir, sinó que "ho va escriure i més d'una vegada i estava molt pensat".
```

Fixem-nos que el sistema ha comès alguns errors, com per exemple classificar com a organització a Maragall.

2.1.1. Estratègies per al reconeixement d'entitats amb nom

Com en moltes altres aplicacions de processament del llenguatge natural, els sistemes de reconeixement d'entitats amb nom es poden classificar en sistemes que funcionen amb regles creades manualment i els sistemes que funcionen a partir d'aprenentatge automàtic a partir de corpus anotats.

Sistemes amb regles creades manualment

Aquests sistemes funcionen amb regles creades manualment. Aquestes regles normalment són algun tipus d'expressió regular.

> Una regla de l'estil "Sr. més una o més Paraula_Capitalizada" podria servir per a detectar noms de persones. Per a trobar noms d'empreses es podria fer servir una expressió de l'estil "una o més Paraula_Capitalizada + SA".

En el programa següent en Python (simple-ner.py) trobem la implementació d'aquestes expressions regulars. Python permet optimitzar molt més la declaració d'expressions regulars, però per claredat mostrem aquest programa tan simple:

```
# -*- coding: utf-8 -*-

import re

expreg1=re.compile('Sr\. ([A-Z][a-z]+ [A-Z][a-z]+ [A-Z][a-z]+) [a-z]')
expreg2=re.compile('Sr\. ([A-Z][a-z]+ [A-Z][a-z]+) [a-z]')
```

```
expreg3=re.compile('Sr\. ([A-Z][a-z]+) [a-z]')

expreg4=re.compile('([A-Z][a-z]+ [A-Z][a-z]+ S\.[AL]\.)')

text="El Sr. Antonio Martinez Gonzalez ha comprat l'empresa Hermanos Saura S.A.
i el Sr. Luis Gomez ha comprat
accions de l'empresa Transportes Lopez S.L. El Sr. Gomez està satisfet amb la seva
nova adquisició."

trobats=expreg1.findall(text)
print "PERSONA:",trobats
trobats=expreg2.findall(text)
print "PERSONA:",trobats
trobats=expreg3.findall(text)
print "PERSONA:",trobats
trobats=expreg4.findall(text)
print "EMPRESA:",trobats
```

Aquest programa donaria la sortida següent:

```
PERSONA: ['Antonio Martinez Gonzalez']
PERSONA: ['Luis Gomez']
PERSONA: ['Gomez']
EMPRESA: ['Hermanos Saura S.A.', 'Transportes Lopez S.L.']
```

Per a poder construir un extractor complet hauríem de crear moltes més regles d'aquest estil. Les regles, a més de detectar algun tipus d'entitat amb nom haurien de disposar de la informació suficient per a poder classificar-les correctament. A més el classificador hauria de combinar-se amb un programa que separés el text en paraules (*tokenitzador*) i que intentés, començant per cada paraula del text, trobar una expressió regular de la llista que coincidís amb el text d'entrada. Si alguna d'aquestes expressions coincideix, es classifica la seqüència de paraules i continua el procés amb la paraula següent després de l'entitat trobada. És possible que diverses expressions regulars coincideixin en algun punt del text, llavors s'haurà de crear algun tipus d'heurística que permeti triar-ne la més adequada. Sovint es tria la seqüència més llarga o es fa servir informació de prioritat associada a les regles.

Aquests tipus de sistemes normalment fan servir llistes predefinides d'entitats conegudes amb la seva classificació: noms de persona, d'empreses, etc.

Tot i que en aquests sistemes les regles es creen manualment, és útil disposar d'un corpus etiquetat per a anar provant el sistema a mesura que es desenvolupen les regles.

Tècniques d'aprenentatge automàtic aplicades a reconeixement d'entitats amb nom

En el camp del reconeixement d'entitats amb nom s'han aplicat un gran nombre de tècniques d'aprenentatge automàtic. A Nadeau i Sekine (2009) podem trobar una bona recopilació d'algorismes d'aprenentatge automàtic aplicats a la tasca de reconeixement d'entitats amb nom. Aquests algorismes es poden classificar en tres grans grups:

1) Aprenentatge supervisat

2) Aprenentatge semisupervisat

3) Aprenentatge no supervisat

Aprenentatge supervisat

La idea és disposar d'un corpus d'aprenentatge en el qual les entitats amb nom estan prèviament marcades i classificades. A partir d'una porció d'aquest corpus s'entrena un sistema i es fa servir la resta del corpus per a avaluar els resultats. A continuació podem observar un fragment del corpus d'aprenentatge per al castellà CoNLL-2002 que es distribueix amb l'NLTK.

```
Por SP O
su DP O
parte NC O
, Fc O
el DA O
Abogado NC B-PER
General AQ I-PER
de SP O
Victoria NC B-LOC
, Fc O
Rob NC B-PER
Hulls AQ I-PER
, Fc O
indic VMI O
que CS O
no RN O
hay VAI O
nadie PI O
que PR O
controle VMS O
que CS O
las DA O
informaciones NC O
contenidas AQ O
en SP O
```

```
CrimeNet NC B-MISC
son VSI O
veraces AQ O
. Fp O
```

El corpus inclou informació sobre la categoria gramatical de les paraules i sobre les entitats amb nom propi. Aquestes estan marcades amb B (que indica inici) i I (que indica interior). En aquest fragment trobem marcats com a persones "Abogado General", "Rob Hulls"; com a localitat "Victoria" i com a altres "Crimenet".

Les tècniques d'aprenentatge automàtic aplicades al reconeixement d'entitats amb nom són diverses:

- Models Ocults de Markov (*Hidden Markov Models* –HMM) (Bikel i altres, 1997).
- Arbres de decisió (*Decision Trees*) (Sekine, 1998).
- Models de màxima entropia (*Maximum Entropy Models* –ME) (Borthwick i altres, 1998).
- Màquines de vectors de suport (*Support Vector Machines* –SVM) (Asahara i Matsumoto, 2003).
- *Conditional Random Fields* –CRF– (McCallum i Li, 2003).

A Ferrández i altres (2005) es presenta un sistema que divideix el problema en dues tasques: detecció i classificació. Per a les dues tasques fan servir tres algorismes d'aprenentatge automàtic: models Ocults de Markov, models de màxima entropia i aprenentatge basat en memòria (*Memory-Based Learning*).

Aprenentatge semisupervisat

La tècnica que més es fa servir per a l'aprenentatge semisupervisat és l'anomenat *bootstrapping*, que parteix d'unes poques dades inicials (en forma de corpus de mida petita o d'exemples de resultats) per a entrenar un sistema inicial. Aquest sistema es fa servir per a etiquetar un corpus més gran o per a obtenir més exemples. A partir dels millors casos trobats es torna a entrenar un sistema. El procés es repeteix fins que es considera que s'ha assolit una precisió i una cobertura suficients o fins que el sistema ja no millora més. A Nadeau i altres (2006) es descriuen uns experiments sobre tècniques semisupervisades que assoleixen uns resultats que s'apropen als dels sistemes que fan servir tècniques d'aprenentatge supervisat.

A Carreras i altres (2003) s'entrena un sistema per al català sense fer servir recursos etiquetats per aquesta llengua. El sistema fa servir tècniques d'aprenentatge automàtic utilitzant recursos per al castellà. Després fan servir dues aproximacions: entrenar models per al castellà i després traduir aquests models al català

o bé entrenar directament models bilingües. Un cop es tenen els primers models s'etiqueten corpus per al català i es fa servir *bootstrapping* per a millorar el sistema.

Aprenentatge no supervisat

La tècnica més habitual per a l'aprenentatge no supervisat d'entitats amb nom és mitjançant clusterització basat en similituds de context (Lin i Pantel, 2001). Altres tècniques es basen en recursos lèxics, com per exemple el Wordnet (Alfonseca i Manandhar, 2002). A Shinyama i Sekine (2004) es basen en l'observació que les entitats amb nom tendeixen a aparèixer simultàniament en diversos articles de notícies d'un mateix dia. D'aquesta manera obtenen llistes que serveixen per a fer-les servir en altres sistemes.

2.2. Resolució de la coreferència

La **resolució de la coreferència** (en anglès, *coreference resolution,* CO) és el procés de determinar si dues expressions en llenguatge natural es refereixen a la mateixa entitat.

Aquestes entitats es refereixen a les detectades pels sistemes de reconeixement d'entitats amb nom i a les seves referències anafòriques.

El següent és un exemple adaptat de Martí (2004):

Barack Obama visitarà el nostre país aquesta setmana. El president dels Estats Units es reunirà amb els màxims dignataris europeus durant el seu viatge. Obama no ha volgut fer cap declaració prèvia.

En aquest exemple podem trobar tres tipus de coreferències:

1) Les de dues entitats amb nom propi que es refereixen a la mateixa entitat (*Barack Obama* i *Obama*)

2) Les que involucren referències pronominals o anafòriques (*el seu viatge* i *Obama*)

3) Les que impliquen coneixement del món (*president dels Estats Units* i *Barack Obama*)

Aquesta tasca no presenta un interès per si mateixa per a l'usuari final, però serveix de tasca auxiliar per a altres tasques relacionades amb l'extracció d'informació.

Lectures complementàries

No entrarem en detalls sobre les metodologies per a portar a terme aquesta tasca. Si voleu ampliar els vostres coneixements podeu consultar:

M. A. Martí (2004). *Tecnologías del texto y del habla* (núm. 72 de la Col·lecció UB). Edicions Universitat de Barcelona.

S. M. Weiss (2005). *Text mining: predictive methods for analyzing unstructured information*. Springer.

2.3. Extracció d'esdeveniments

L'**extracció d'esdeveniments** (en anglès, *Event Extraction*) és una tasca que consisteix a trobar tots els casos d'un determinat tipus de relació o d'esdeveniment d'un text o conjunt de textos.

Per exemple, d'un text com el següent:

Caixa Terrassa presenta 35,8 milions de guany el 2009 i ultima la fusió d'Unnim. La nova entitat resultant de la integració amb Sabadell, Girona i Manlleu podria començar a operar en els mercats a mitjan juny.

Caixa Terrassa ha presentat aquest divendres el resultat de l'exercici de 2009, que ha tancat amb un guany de 35,8 milions d'euros, un 27,6% menys després de destinar 113,4 milions a provisions per a insolvències, i consolida el procés de fusió amb les caixes de Sabadell, Girona i Manlleu. La integració "compleix amb tots els requisits de Brussel·les", segons el director general de l'entitat, Enric Mata, que serà també el màxim directiu aquestes entrades, d'Unnim, l'entitat resultant de la fusió de les quatre caixes.

Es podrien generar les entrades[1] següents d'una base de dades:

(1)Sovint a aquestes entrades se les anomena *plantilles*.

```
Resultat empresa:
Expresa: Caixa Terrassa
Exercici: 2009
Resultat: 35.8 M
```

i també:

```
Fusions empresa:
Resultant: Unnim
Fusionades: Caixa Terrassa, Caixa Sabadell, Caixa Girona, Caixa Manlleu
Inici activitat: juny-2010
```

Així, doncs, l'objectiu d'aquesta tasca és convertir informació no estructurada que es troba en textos en informació estructurada. Aquesta tasca és complexa i molt depenent del domini. La tasca s'acostuma a portar a terme mitjançant un conjunt de patrons que permeten detectar esdeveniments d'un determinat tipus. Aquests patrons es poden desenvolupar manualment o bé crear-los automàticament. A Turmo i altres (1998) es presenta un sistema que pot construir esquemes sintacticosemàntics i regles d'extracció rellevants per a un domini a partir d'un corpus d'aprenentatge i una ontologia per a aquest domini.

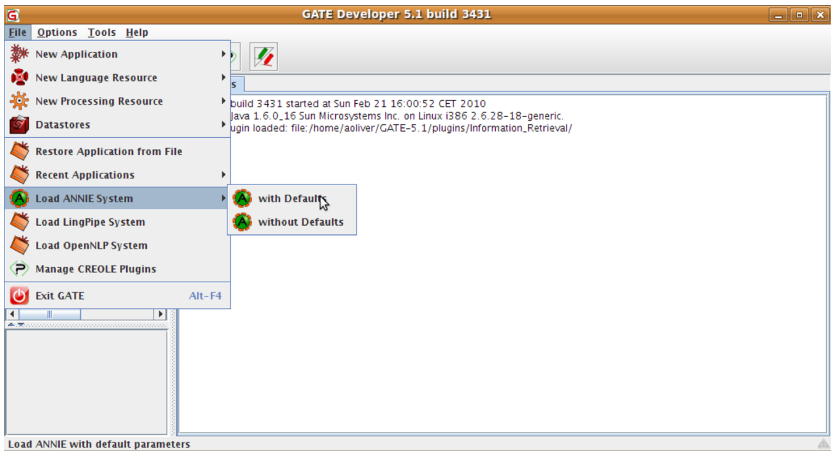
A Black i Ranjan (2004) es presenta un sistema que pretén detectar informació d'esdeveniments a partir dels missatges de correu electrònic que rep un usuari. Un dels objectius és poder construir l'entrada d'un determinat esdeveniment per a poder-la incloure al calendari de l'usuari.

2.4. GATE com a sistema d'extracció d'informació

En el subapartat 1.3. vam presentar GATE i vam aprendre a aplicar-lo a la tasca de recuperació d'informació. En aquest subapartat veurem l'aplicació d'aquesta eina a una tasca relacionada amb l'extracció d'informació.

GATE disposa d'un sistema complet d'extracció d'informació que s'anomena *a Nearly-New Information Extraction System* (ANNIE). Per a fer servir ANNIE hem de seleccionar l'opció *Load ANNIE System* del menú *File* (figura 14). Per a fer-lo servir amb les seves opcions per defecte escollirem l'opció *With Defaults*. D'aquesta manera carregarem tots els recursos que necessita ANNIE i crearem un corpus *Pipeline* que s'anomenarà ANNIE amb els recursos correctes seleccionats en l'ordre correcte i amb els conjunts d'anotació d'entrada i sortida per defecte. Si es selecciona l'opció *Without Defaults* es carregaran els mateixos recursos de processament però per a cada recurs apareixerà un diàleg que permetrà a l'usuari seleccionar la ubicació i el nom del recurs.

Figura 14. Càrrega d'ANNIE a GATE



Per a executar ANNIE necessitarem també tenir carregat el corpus que volem tractar. Per a fer-ho hem de fer *File* > *New Language Resource* > *GATE Corpus* (figura 15). Un cop creat el corpus el seleccionem en la pantalla de l'esquerra i fem clic al botó dret del ratolí i seleccionem *Populate*. Això ens permet carregar tots els arxius que hi hagi en un determinat directori.

Ara ja ho tenim tot per a poder executar ANNIE. A la pantalla de l'esquerra busquem *Applications* i fem doble clic a ANNIE. Apareixerà una pantalla com la de la figura 16. En aquesta pantalla podem seleccionar el corpus sobre el que volem executar ANNIE (que ha de ser algun dels corpus carregats a sistema). Per a executar ANNIE hem de fer clic al botó *Run this Application*.

Vegeu també

En el subapartat següent veurem el resultat de l'execució d'ANNIE. Concretament veurem els resultats de la detecció d'entitats amb nom.

Figura 15. Creació d'un corpus en GATE

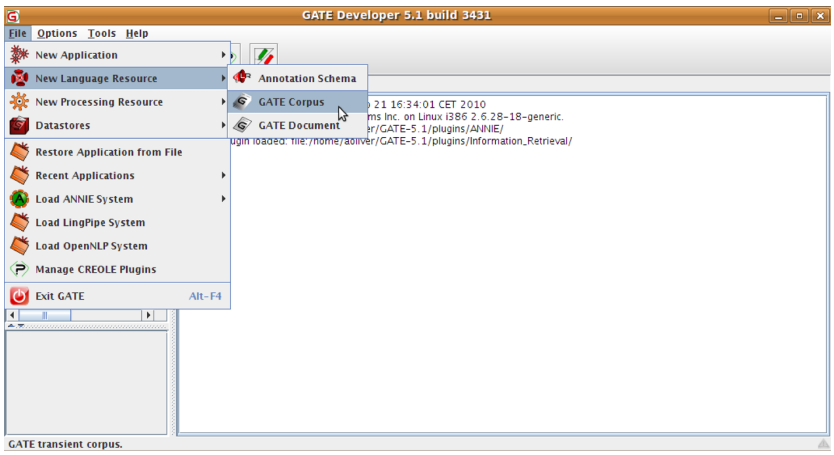
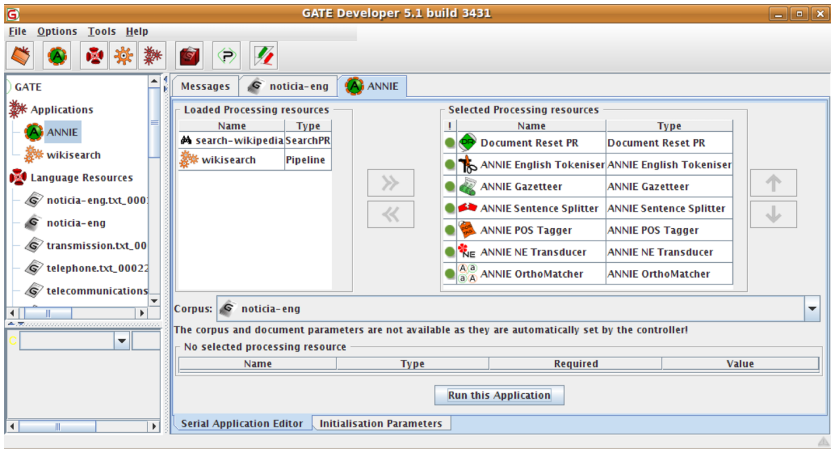


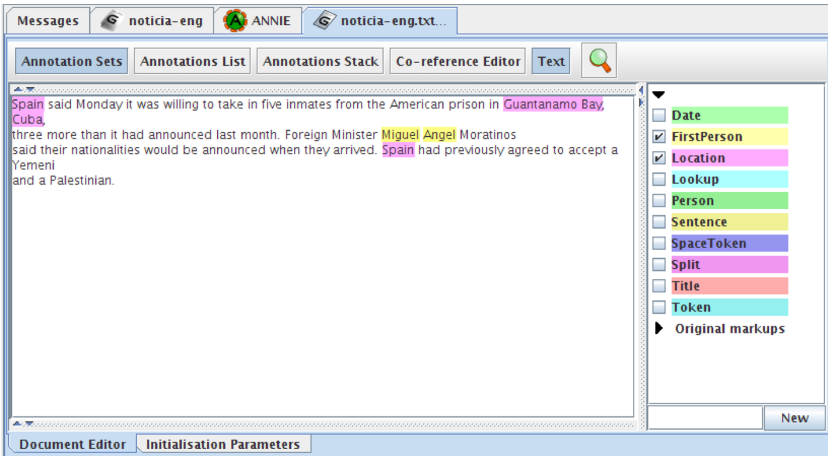
Figura 16. Execució d'ANNIE



2.4.1. Detecció d'entitats amb nom amb GATE

Per a veure els resultats hem de fer doble clic sobre l'arxiu processat per ANNIE a la pantalla de l'esquerra i fer clic al botó *Annotation Sets* per a veure els diferents tipus d'entitats que detecta. Si marquem algun d'aquests tipus es marcaran les entitats corresponents al text (figura 17).

Figura 17. Detecció d'entitats amb nom amb GATE



3. Sistemes de pregunta-resposta

Fins ara en aquest mòdul hem vist els sistemes de recuperació d'informació (*information retrieval*) i els d'extracció d'informació (*information extraction*). Els sistemes de recuperació d'informació pretenen retornar un seguit de documents que siguin rellevants respecte a una consulta formulada per l'usuari. Els sistemes d'extracció d'informació són més heterogenis i es poden classificar segons diverses tasques. Una d'aquestes tasques és prendre certa informació que està de manera no estructurada en els textos i convertir-la en informació estructurada, com a entrades d'una base de dades, per exemple.

Els sistemes de pregunta-resposta (en anglès, *question-answering*, QA) pretenen retornar la resposta concreta a una pregunta de l'usuari basant-se en el contingut d'un conjunt de textos.

> Els sistemes de pregunta-resposta són unes aplicacions de processament del llenguatge natural en les quals, atesa una pregunta formulada en llenguatge natural i una col·lecció de documents, retornen una resposta exacta a la pregunta partint de la informació que hi ha a la col·lecció de documents (Hermjakob, 2006).

Per a trobar les respostes els sistemes de QA poden fer servir tant una col·lecció de documents en llenguatge natural (una col·lecció pròpia de documents, la Web, la Viquipèdia, etc.) o bé una base de dades d'informació estructurada (com les que hem vist a l'apartat 2, dedicat als sistemes d'extracció d'informació).

Aquests sistemes han de ser capaços de respondre a molts tipus de preguntes diferents:

- fets: per exemple, *Qui va guanyar el premi Nobel de literatura el 1936?*
- llistes: per exemple, *Quines són les principals ciutats d'Espanya?*
- definicions: per exemple, *Què és una base de dades?*
- preguntes tipus Com?: per exemple *Com es fabrica una bombeta?*
- preguntes tipus Per què?: per exemple, *Per què Alemanya va envair Polònia?*

A més, alguns sistemes de QA resolen altres qüestions, com ara *Converteix-me 10 euros en dòlars* o fins i tot preguntes de l'estil *Quin país és més gran, Alemanya o França?*

Els sistemes de QA es poden classificar segons el domini:

1) Sistemes per a un domini específic: pretenen respondre a preguntes restringides a una àrea de coneixement específica (medicina, dret, manteniment i reparació d'automòbils, etc.).

2) Sistemes de domini obert: pretenen respondre a preguntes sobre qualsevol tema.

3.1. Arquitectura dels sistemes de QA

En general, els sistemes de QA disposen de tres mòduls:

- Mòdul de processament de les preguntes
- Mòdul de processament dels documents
- Mòdul d'extracció de la resposta

3.1.1. Mòdul de processament de les preguntes

Aquest mòdul implementa una sèrie de tècniques que permeten la interpretació de les preguntes per a poder trobar posteriorment les respostes.

Sistemes de QA i llenguatge natural

Recordeu que a diferència dels sistemes de recuperació d'informació, on les preguntes es formulen mitjançant una sèrie de paraules clau acompanyades o no d'alguns operadors (AND, OR...), en els sistemes de QA les preguntes es formulen en llenguatge natural.

Una de les tasques principals d'aquest mòdul és determinar el tipus de resposta esperada. Per exemple, d'una pregunta com *Qui va descobrir Amèrica?* el tipus esperat és *PERSONA*. Una altra tasca important d'aquest mòdul és determinar quines són les paraules de la pregunta que es poden fer servir per a consultar els índexs dels documents per trobar els paràgrafs dels documents on es pot trobar la resposta.

3.1.2. Mòdul de processament dels documents

Aquest mòdul és l'encarregat d'indexar els documents de la col·lecció d'alguna manera que serveixi per recuperar els paràgrafs dels documents on es poden trobar les respostes. Alguns sistemes requereixen que els paràgrafs recuperats continguin totes les paraules claus de la pregunta i com a mínim una paraula de la mateixa categoria semàntica del tipus de resposta esperat (Moldovan i altres, 2000; Clarke i altres, 2000).

A Ittycheriah i altres (2001) es presenta un sistema que processa els documents en dues passades. A la primera passada busquen en una base de dades enciclopèdica. Els passatges que obtenen una puntuació més alta es fan servir per a expandir les preguntes i fer una segona passada sobre la col·lecció de documents.

3.1.3. Mòdul d'extracció la resposta

L'extracció de la resposta és el procés d'identificar en un paràgraf i fragment de text que representa la resposta a la pregunta formulada.

Un aspecte important a tenir en compte és la llargada de la resposta que s'ofereix a l'usuari. L'ideal seria una resposta curta i precisa. Si la resposta que retorna el sistema és molt curta és possible que no es mostri la informació desitjada i si és massa llarga és possible que ofereixi informació no desitjada.

3.2. Conclusions

Hem presentat els principis bàsics dels sistemes de pregunta-resposta. Per a ampliar coneixements podeu consultar a Vicedo i altres (2003). En aquest article es presenta una classificació detallada dels diferents sistemes de QA. També es presenten dos sistemes: el SEMQA, desenvolupat a la Universitat d'Alacant i el sistema multilingüe de la UdG i la UPC. També dedica una secció a parlar sobre l'avaluació d'aquest tipus de sistemes.

3.3. El sistema QA START

En aquest subapartat veurem en funcionament un sistema de QA que està disponible a Internet, l'Start. Aquest és un dels primers sistemes de QA que es basen en la informació de la Web i està en funcionament des de desembre de 1993. Va ser desenvolupat per l'equip de Boris Katz del Gruo InfoLab del MIT Computer Science and Artificial Intelligence Laboratory.

Web recomanat

El sistema QA Start està disponible a http://start.csail.mit.edu/. Des d'aquesta web es pot accedir al sistema i trobar informació addicional sobre el seu funcionament. Aquest sistema funciona només per a l'anglès.

3.3.1. Funcionament bàsic

El sistema Start[2] analitza sintàcticament les preguntes que fan els usuaris i fa servir aquesta anàlisi sintàctica per a cercar en la seva base de coneixement. Aquest sistema fa servir una tècnica que s'anomena *anotació en llenguatge natural* (*natural language annotation*). Aquesta tècnica fa servir anotacions en llenguatge natural com a descriptors del contingut que s'associen a segments d'informació. Un segment d'informació se selecciona quan les seves anotacions coincideixen amb la pregunta de l'usuari. Això permet al sistema recuperar no només text, sinó també contingut multimèdia, ja que ha estat anotat.

[2]Start són les sigles de *SynTactic Analysis using Reversible Transformations.*

El component de processament del llenguatge natural d'Start té dos mòduls que comparteixen la mateixa gramàtica. El mòdul de comprensió analitza el text i produeix una base de coneixement que codifica la informació que hi ha

en el text. El mòdul de generació produeix frases en anglès a partir del segment apropiat de la base de coneixement. El segment de la base de dades de coneixement s'haurà trobat a partir de la tècnica d'anotació explicada anteriorment.

Activitat

Proveu aquest sistema amb diferents tipus de preguntes. A la mateixa pàgina web trobareu un bon grapat d'exemples.

4. Recuperació d'informació multilingüe

En els apartats anteriors hem vist tècniques de recuperació d'informació, d'extracció d'informació i de pregunta-resposta que estan dissenyades per a treballar amb una única llengua (i en la majoria dels casos l'anglès).

En aquest apartat presentarem els principals problemes i tècniques relacionats amb la recuperació d'informació multilingüe. També presentarem breument els avenços en relació amb l'extracció d'informació i amb els sistemes de pregunta resposta multilingües.

Tots aquests sistemes intenten explotar la informació dels documents que estan en més d'una llengua i respondre a peticions d'informació que poden estar expressades en més d'una llengua.

4.1. Aproximacions per a la recuperació d'informació multilingüe

Hi ha diverses aproximacions al problema de la recuperació d'informació multilingüe. Se'n poden destacar les següents (les primeres extretes de Fluhr (2000)):

- Ús d'un vocabulari controlat per a la indexació i la recuperació.
- Ús de sistemes de traducció automàtica.
- Ús del model d'espai vectorial de Salton.
- Ús de corpus multilingües.

4.1.1. Ús d'un vocabulari controlat per a la indexació i la recuperació

Una de les aproximacions tradicionals a la recuperació d'informació en general i a la multilingüe en particular es basa en l'ús d'un vocabulari controlat tant per a indexar els documents com per a fer les consultes. Un documentalista o bé un programa informàtic selecciona per a cada document de la col·lecció uns descriptors d'una llista tancada de termes autoritzats. Aquesta llista s'organitza en forma de tesaurus de manera que es disposa de les relacions semàntiques entre els termes. Per a poder portar a terme la recuperació multilingüe cada terme del tesaurus es tradueix en cada una de les llengües admeses al sistema. Aquesta aproximació dóna uns resultats acceptables però no permet fer consultes que no estiguin relacionades amb els termes permesos.

A Soergel (1997) podem trobar molts més detalls sobre aquesta metodologia, exemples de tesaurus multilingües i sistemes reals que implementen aquesta aproximació.

4.1.2. Ús de sistemes de traducció automàtica

Alguns sistemes fan servir sistemes de traducció automàtica per a traduir les consultes i fins i tot tota la col·lecció de documents. En els casos en què només es tradueix la consulta, els documents que es recuperen en la llengua d'arribada es tradueixen automàticament de manera dinàmica a la llengua de partida de la consulta. El problema amb aquest tipus de sistemes és que els errors de traducció afecten els resultats de recuperació.

4.1.3. Ús del model d'espai vectorial de Salton

L'espai vectorial de Salton (Salton i altres, 1975) representa els documents en un espai d'n dimensions, en el qual n és el nombre de paraules diferents en la col·lecció de documents. Si alguns documents estan traduïts a una altra llengua, aquests documents es poden representar tant en el subespai relacionat amb la primera llengua com en el subespai relacionat amb la segona. Si el sistema rep una consulta expressada en la segona llengua, es poden obtenir els documents més rellevants fent servir alguna mesura de proximitat. Aquests documents recuperats es poden fer servir per a extreure documents semblants en el subespai de la primera llengua.

4.1.4. Ús de corpus multilingües

A Braschler i Schäuble (2000) es presenta un sistema de recuperació d'informació multilingüe que fa servir un corpus multilingüe per a construir estructures de dades de manera automàtica que permeten traduir la consulta, unir els resultats de la recuperació per a les diferents llengües i assignar la rellevància de cada document independentment de la llengua en què estigui escrit.

La col·lecció de documents multilingües s'alinien a nivell de document. És a dir, es relacionen els documents en dues llengües o més que estan molt relacionats, tot i que no necessàriament han de ser la traducció l'un de l'altre. Aquesta relació entre documents es fan mitjançat **indicadors**. Aquests indicadors poden ser noms propis i xifres o dates compartits entre documents i termes que es puguin traduir a partir d'un diccionari bilingüe.

4.2. Extracció d'informació multilingüe

A Gaizauskas i altres (1997) es presenten tres alternatives per a portar a terme la tasca d'extracció d'esdeveniments en un entorn en el qual intervinguin dues llengües (llengua A i B):

Vegeu també

Sobre l'extracció d'esdeveniments vegeu el subapartat 2.3.

- Si disposem d'un sistema de traducció automàtica de B a A i d'un sistema d'extracció d'informació de la llengua A, podem traduir tots els documents de la llengua B a la llengua A i fer servir el sistema d'extracció d'informació de la llengua A.

- Disposem d'un sistema d'extracció d'informació per a la llengua A i d'un altre per a la llengua B i els fem servir per a extreure les entrades de la base de dades o **plantilles**. Després un sistema de traducció automàtica "mini" traduirà les entrades de la base de dades de la llengua B a la llengua A. Anomenen a aquest sistema de traducció automàtica "mini" perquè les entrades o plantilles tenen una informació molt estructurada i bàsicament són entrades lèxiques. Moltes d'aquestes entrades lèxiques són noms propis que no cal traduir.

- Fer servir un model del domini que sigui independent de la llengua. En aquest model els conceptes estan relacionats a entrades lèxiques en diverses llengües mitjançant diccionaris. Aquest model del domini es fa servir per a produir representacions dels textos de la col·lecció que siguin independents de la llengua, i ho anomenen *model del discurs*. A partir d'aquest model del discurs es produeixen les entrades o plantilles que contenen la informació. La informació d'aquestes plantilles es pot presentar en qualsevol de les llengües del sistema.

La resta de l'article que hem citat la dedica a fer una proposta concreta de la tercera de les possibilitats.

4.3. Sistemes de pregunta-resposta multilingües

Alguns intents de desenvolupament de sistemes de pregunta-resposta multilingüe es basen en la traducció automàtica de les preguntes i de la col·lecció de documents. A Shukla i altres (2004) es presenta una proposta que converteix tant les preguntes com els documents de la col·lecció a una representació intermèdia fent servir una interlingua. Concretament fa servir el *Universal Networking Language* (UNL), que és un llenguatge artificial que es pot fer servir com a llengua pivot en els sistemes de traducció automàtica basats en interlingua o com a llenguatge de representació del coneixement en sistemes de recuperació i extracció d'informació.

5. Els cercadors d'Internet

En aquest apartat veurem uns sistemes relacionats amb la recuperació d'informació que fan servir com a conjunt de documents una fracció del contingut d'Internet. De tots els sistemes que hem vist en aquest mòdul és, sens dubte, el que ha assolit un grau d'utilització més gran en tots els tipus d'usuaris. Qui no ha intentat cercar alguna cosa a Google, Yahoo, Altavista o d'altres? Qui no fa servir algun servei associat a un cercador d'Internet (correu electrònic, per exemple)?

Els cercadors d'Internet tenen com a principal objectiu (i a la vegada principal dificultat) indexar "tot" el contingut d'Internet (bé, tant com sigui possible) i oferir un conjunt d'enllaços a documents rellevants a la cerca de l'usuari endreçats per rellevància amb uns temps de resposta molt ràpids.

De fet, no es coneix la mida exacta d'Internet. Penseu que Internet és molt dinàmica, hi apareixen nous continguts,hi desapareixen d'altres. Part del contingut està en forma de bases de dades, cosa que dificulta la seva indexació pels motors de cerca.

Hi ha algunes pàgines web que intenten valorar la mida de la web indexada pels principals cercadors. Per exemple, a http://www.worldwidewebsize.com/ podem observar gràfics detallats de l'evolució de les pàgines indexades. Aquest web estima que, el 10 de març de 2010, hi ha més de 19.000 milions de pàgines web indexades.

Els primers cercadors eren llistes de pàgines web endreçades temàticament. Aquestes llistes al principi es feien manualment, és a dir, hi havia un equip de persones que visitava les pàgines web i les classificava segons una jerarquia de categories i hi assignava una sèrie d'etiquetes.

Posteriorment apareixen els primers programes informàtics que visiten automàticament les pàgines web i les indexen segons el seu contingut. També apareixen els primers cercadors que es basen en aquests programes i utilitzen els índexs creats per a oferir una llista d'enllaços endreçats per rellevància. Els primers cercadors basaven la rellevància en el contingut de la pàgina web i en les paraules que el creador de la web posava sota l'etiqueta "meta" d'HTML. Això feia que els mateixos creadors de les pàgines web poguessin manipular de manera més o menys fàcil el funcionament dels cercadors i que els resultats de les cerques no retornés documents realment rellevants.

L'algorisme conegut com a PageRank™, que és el que fa servir Google per a endreçar els seus resultats, fa servir una idea totalment nova i que situa Google entre els cercadors més emprats.

La lluita entre els principals cercadors d'Internet (principalment Google i Yahoo) fa que ofereixin molts més serveis (correu electrònic, cerca de vídeos o música, cerca a mapes i càlcul de rutes i un llarg etcètera) amb l'objectiu de captar el màxim nombre d'usuaris.

Model de negoci dels cercadors

Sobre el model de negoci d'aquests cercadors fa uns anys es rumorejava que Google faria pagar diners per a poder fer cerques. La maniobra d'aquest cercador i d'altres va ser no fer pagar als usuaris, sinó basar el model de negoci principalment en la publicitat.

D'altra banda els principals cercadors estan apostant no només per indexar millor, oferir millors índexs de rellevància i indexar el màxim nombre de pàgines, si no per explorar noves formes de mostrar i interrelacionar la informació.

En aquest apartat veurem una mica tots aquests aspectes relacionats amb els cercadors d'Internet. Farem molta referència a productes principalment de Google i també de Yahoo. La tria no respon a cap interès especial de l'autor, si no que simplement en el moment de redacció d'aquest mòdul aquestes dues empreses eren les més importants.

5.1. *Open Directory Project*

Com ja hem comentat a la introducció d'aquest mòdul, els primers cercadors d'Internet es basaven en llistes de pàgines web classificades temàticament. Aquestes llistes es confeccionaven manualment i un equip humà les visitava, les classificava i hi assignava una sèrie de descriptors. Actualment existeix un projecte col·laboratiu anomenat *Open Directory Project* que fa aquesta tasca manual amb l'ajut d'editors voluntaris. S'hi pot accedir en diversos idiomes, entre ells el català i el castellà. A més, com que és un projecte lliure, les dades es poden descarregar.

Les dades d'aquest projecte s'aprofiten en altres cercadors, com ara Google, que disposa d'una interfície de cerca igual que la de l'Open Directory Project. Google afegeix només la informació de la rellevància de cada una d'aquestes pàgines a partir dels seus algorismes.

Webs recomanades

Es pot accedir al web de l'Open Directory Project a http://www.dmoz.org/.

També podeu accedir a la interfície de Google Directory a http://www.google.com/dirhp.

5.2. L'algorisme Page Rank™ de Google

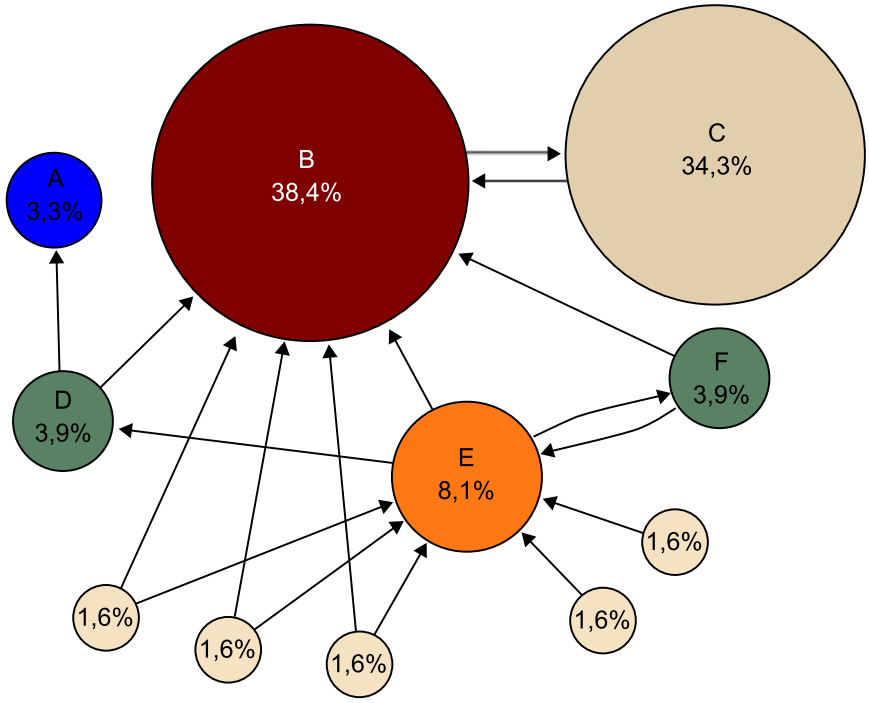
Una de les claus de l'èxit del cercador Google ha estat l'algorisme que fa servir per a calcular la rellevància de les pàgines que mostra. Els primers cercadors feien servir simplement l'aparició de les paraules de la cerca als documents

i com més vegades apareguessin les paraules i en llocs determinats del document (títols, etiqueta meta, etc.) millor. Això feia que els creadors de la pàgines web poguessin manipular fàcilment els resultats de la rellevància.

Google fa un càlcul de rellevància que es basa principalment en l'algorisme Page Rank™. Aquest algorisme calcula una mesura objectiva de la rellevància que tenen les pàgines web dins la Xarxa i es basa a assignar un valor a cada web en funció del nombre d'enllaços d'altres pàgines que l'apunten, interpretant un vincle de la pàgina A a la pàgina B com un vot que rep la pàgina B per part de la pàgina A. A més, PageRank també considera el prestigi de cada pàgina que emet un vot, ja que als vots que provenen de determinades pàgines se'ls atorga un valor major, incrementant així el valor de la pàgina vinculada. D'aquesta manera i juntament amb altres criteris, les pàgines importants reben una valoració més alta i apareixen en la part superior dels resultats de cerca.

A la figura 18 es mostra gràficament el funcionament d'aquest algorisme. A la figura observem que la pàgina més rellevant és la B ja que hi ha moltes pàgines que l'enllacen. La C està en segona posició tot i que només té una pàgina que l'enllaça, però és una pàgina molt important. La E rep molts enllaços, però de pàgines poc importants.

Figura 18. Representació gràfica de l'algorisme Page Rank™ de Google



Font: Vikipèdia.

PageRank és la part més coneguda de l'algorisme de classificació de Google, però també fa servir altres tècniques, com per exemple:

- Models de llenguatge: que modelen sinònims als termes de cerca, errors ortogràfics, etc.

- Models de consulta: que modelen la manera com els usuaris formulen les consultes.
- Models temporals: que tenen en compte la data de creació de les pàgines web i donen una mica més d'importància a pàgines més recents.
- Models personalitzats: que tenen en compte com els usuaris concrets (si estan donats d'alta al sistema) fan les consultes.

Aquests models es confeccionen de manera pràcticament automàtica tenint en compte les consultes dels usuaris i els clics que fan a les pàgines que proposa el cercador. Tota intervenció de l'usuari s'emmagatzema i s'analitza de manera automàtica per a millorar el funcionament del cercador.

5.3. Funcionalitat de cerca avançada de Google

Si feu servir un cercador qualsevol, ja sigui Google, Yahoo o d'altres és molt important que conegueu quines són les possibilitats de cerca avançada. Aquestes possibilitats ens ajudaran enormement a refinar les nostres cerques i a trobar més fàcilment allò que volem. Per a tenir accés a totes les possibilitats de cerca convé tenir seleccionada la interfície en anglès, ja que les darreres novetats triguen una mica a estar a les versions localitzades en altres llengües. Per a accedir a les opcions avançades hi ha un enllaç a prop del botó de cerca. En el moment d'escriure aquests materials Google oferia les següents possibilitats de cerca avançada en la seva interfície en anglès:

- Pàgines que tenen totes o alguna de les paraules de la cerca, la frase exacta, que no tingui cap paraula de la llista.
- Pàgines escrites en una determinada llengua.
- Fitxers en un format determinat (PDF, DOC, RTF, etc.).
- Les pàgines d'un domini o un lloc web específic (per exemple, cercar només dins de les pàgines de www.uoc.edu).
- En una data determinada (les darreres 24 hores, la darrera setmana, el darrer mes, etc.).
- Segons la llicència d'ús, és a dir, contingut que es pot modificar o no, distribuir, fer servir de manera comercial, etc.
- Que les paraules de la cerca estiguin en un lloc determinat, títol, cos, en la URL, etc.
- Restringir la regió o país on està publicada la pàgina.
- Especificar un rang numèric.
- Pàgines similars a una pàgina determinada.
- Pàgines que enllacen a una altra pàgina.

Algunes de les opcions de cerca es poden fer servir directament amb comandes. Per exemple, si volem cercar totes les pàgines que continguin *filologia catalana* (junt, no *filologia* per una banda i *catalana* per una altra) però que no continguin la paraula *literatura* i siguin pàgines de la UOC, podem escriure:

```
"filologia catalana" -literatura site:uoc.edu
```

La interfície de cerca avançada tradueix les seleccions a les ordres corresponents així que podem aprendre la sintaxi fixant-nos en la casella de cerca cada cop que fem servir les opcions avançades.

5.4. Incloure la meva pàgina web en els cercadors

Si disposeu d'una pàgina web que encara no està indexada per algun cercador com Google o Yahoo podeu demanar que us indexin la pàgina. L'adreça que proporcioneu s'inclou a una llista de pàgines que seran indexades. La indexació pot trigar força temps, però s'acabarà indexant.

En el moment d'escriure aquests materials a Google amb la interfície en català s'havia de seguir els enllaços següents: a la pàgina principal fer clic sobre l'enllaç "Tot sobre Google". Després, sota la secció "Per a propietaris de llocs web" fer clic a "Enviar contingut a Google". A partir d'aquest moment la interfície està en anglès i s'ha de fer clic a l'enllaç "Web" i després fer clic a "Add Your URL to Google's Index". Aquí podreu introduir el vostre web, que quedarà a l'espera de ser indexat.

5.5. Alertes de Google

Google disposa d'un servei d'alertes per correu electrònic. Si ens interessa saber si Google indexa un contingut nou que satisfaci uns criteris de cerca podem donar-nos d'alta a aquest servei. Llavors el sistema ens enviarà un correu electrònic a l'adreça indicada. La periodicitat del correu pot ser diversa: quan es produeixi una nova indexació, un cop al dia agrupant totes les noves indexacions o un cop per setmana. Això pot ser interessant per a monitorar algun tema que ens interessi especialment.

Web recomanat

Per a accedir al servei d'alertes de Google cal anar a http://www.google.com/alerts.

5.6. Google Academics

Google disposa d'un servei de cerca específic per a articles acadèmics. Si cerquem aquí buscarà articles relacionats amb la cerca i molt sovint tindrem accés directe a l'article. En les opcions avançades es pot buscar per autor, publicació i dates.

Web recomanat

S'accedeix a Google Academics des de http://scholar.google.com.

Quan tenim la finestra de resultats de cada article que ha trobat disposem de més informació addicional. Per exemple, podem veure un enllaç que indica quants articles el citen, un enllaç a articles relacionats i si l'article disposa de més d'una versió, tenim accés a totes les versions disponibles.

En l'enllaç de preferències de la cerca tenim diverses opcions, entre les qual podem destacar la possibilitat que ens indiqui la referència bibliogràfica de l'article en diversos formats. Això és molt útil per a confeccionar la bibliografia dels nostres treballs.

5.7. Google Books

Google ha fet una aposta molt ambiciosa per a indexar no només el contingut de la Web, sinó també dels llibres publicats arreu del món. Per a assolir aquest objectiu ha arribat a acords amb biblioteques i editorials per a digitalitzar llibres i deixar-los disponibles a Internet i amb les possibilitats de cerca que ofereix Google. Com que molts llibres encara estan protegits per drets d'autor la visualització que s'ofereix és limitada. Això vol dir que no pots visualitzar tot el llibre però sí una bona part d'ell. Això acostuma a ser suficient per a decidir si val la pena comprar el llibre o demanar-lo a alguna biblioteca. Ara bé, els llibres que no tenen restriccions per drets d'autor (perquè disposen d'una llicència d'ús lliure o bé perquè els drets d'autor han caducat) es poden consultar en la seva totalitat i fins i tot descarregar-los en PDF.

Web recomanat

Es pot accedir a Google Books des de http://books.google.com/.

5.8. SandBox de Yahoo

Els grans cercadors d'Internet investiguen molts aspectes relacionats amb la indexació de la informació disponible a Internet i com mostrar la informació més rellevant a l'usuari. Molts aspectes de la investigació actual se centren en com mostrar la informació i com relacionar-la.

El SandBox de Yahoo és un espai de prova de les innovacions de Yahoo. Us invitem a accedir-hi i a veure les darreres innovacions. En el moment de l'escriptura d'aquests apartats es podien destacar els productes següents:

Web recomanat

Es pot accedir a SandBox de Yahoo des de http://sandbox.yahoo.com/

- **Correlator** (http://correlator.sandbox.yahoo.net): extreu i organitza la informació i busca noms, conceptes, llocs i esdeveniments relacionats amb la consulta.

- **Quest** (http://quest.sandbox.yahoo.net/): treballa amb un conjunt de preguntes i respostes i guia l'usuari amb un resum dels termes relacionats més importants de manera que pugui anar restringint la cerca fins a trobar la solució.

- **Motif** (http://motif.sandbox.yahoo.net): és un sistema de recuperació que intenta respondre a partir del context de la cerca.

És interessant visitar aquest lloc de tant en tant per a veure cap a on va la recerca a Yahoo.

5.9. Altres serveis oferts pels principals cercadors

Per a poder atreure el màxim nombre d'usuaris, els principals cercadors d'Internet ofereixen un gran nombre de serveis entre els quals podem destacar:

- Correu electrònic
- Cerca d'imatges, vídeos, música...
- Mapes, cerques d'adreces, càlculs de recorreguts
- Sistemes de traducció automàtica
- Llocs per a crear les vostres pròpies webs, blogs, etc.
- Calendaris i agendes
- Editors de documents en xarxa

I una llarga llista de serveis.

5.10. Finançament del cercadors d'Internet

Els grans cercadors d'Internet ofereixen una sèrie de serveis gratuïts. Com guanyen diners? Una de les grans fonts d'ingressos dels cercadors és la publicitat. Quan feu una cerca us apareixen una sèrie d'enllaços endreçats per rellevància basada en els seus algorismes. Però també ofereixen una sèrie d'*enllaços patrocinats*, és a dir, enllaços pels quals algú ha pagat perquè surtin. De fet, normalment no paguen per sortir, si no que paguen si l'usuari fa clic a l'enllaç. I quant paguen? Doncs depèn de la paraula clau a partir de la qual vol sortir. Hi ha paraules molt cares i paraules més barates. El fet és que és una manera assequible de publicitar els serveis d'una empresa ja que només es paga si un usuari fa clic a l'enllaç. A més l'anunciant pot determinar el màxim que vol pagar en un mes i quan s'assoleix aquesta quantitat l'anunci deixa d'aparèixer.

Ara bé, la política dels cercadors, i molt encertada, és mostrar els enllaços patrocinats de manera que quedi clar que no apareixen perquè tinguin una rellevància gran, sinó perquè han pagat per sortir.

La propaganda va més enllà dels cercadors i Gmail, per exemple, mostra publicitat relacionada amb el missatge de correu que estàs llegint. Això suposa una lectura, ni que sigui automàtica, del contingut d'un missatge privat. I això ha aixecat moltes reticències i un control força exhaustiu per part del govern dels Estats Units. Sigui com sigui, Google explica què fa amb els correus i com els processa i queda a l'elecció de l'usuari acceptar els serveis o no.

La publicitat s'afegeix a la majoria dels serveis que ofereixen, com els mapes, per exemple.

La conclusió és que fem servir uns serveis gratuïts però no els ofereixen de franc. No els paguem però acceptem una publicitat que, si ens interessa mirar, paga algú altre.

5.11. Conclusions

En aquest apartat hem presentat alguns dels conceptes més importants relacionats amb els cercadors d'Internet. Els canvis que es produeixen en aquests serveis són tan ràpids que l'apartat quedarà obsolet en poc temps. Per aquest motiu us convidem a visitar els principals cercadors d'Internet i analitzar les opcions avançades de cerca, així com els serveis que ofereixen.

Resum

En aquest mòdul hem presentat les principals tècniques i aplicacions relacionades amb la cerca i recuperació d'informació. La quantitat d'informació a la qual tenim accés actualment és enorme i és imprescindible disposar d'eines que ens facilitin l'accés.

Els sistemes de recuperació d'informació seleccionen un conjunt de documents d'una col·lecció que són rellevants per a la consulta formulada per l'usuari. Aquests sistemes presenten els documents endreçats per algun índex de rellevància. L'usuari haurà de cercar la informació desitjada en el conjunt de documents retornats.

Els sistemes d'extracció d'informació tenen com a objectiu convertir la informació no estructurada que es troba en una col·lecció de documents en informació estructurada. Els sistemes d'extracció d'informació poden tenir diferents objectius: cercar i classificar entitats amb nom, cercar esdeveniments d'algun tipus, etc.

Els sistemes de pregunta-resposta pretenen contestar de manera precisa a una pregunta formulada en llenguatge natural. A partir de la pregunta ha de trobar la informació i presentar-la en llenguatge natural a l'usuari. La resposta ha de ser precisa i contenir el mínim possible d'informació supèrflua.

La majoria de sistemes que hem presentat han estat dissenyats per a l'anglès i poden funcionar per a altres llengües amb algunes modificacions. També es dissenyen sistemes multilingües capaços de funcionar en diverses llengües simultàniament. La col·lecció de documents i les consultes dels usuaris poden estar en diferents llengües i la informació s'ha de presentar en la llengua triada per l'usuari independentment de la llengua en què estigui el document que conté aquesta informació.

Per acabar hem presentat els aspectes bàsics relacionats amb els cercadors d'Internet, eines imprescindibles per a poder accedir a la informació que hi ha a la Xarxa.

Bibliografia

Alfonseca, E.; Manandhar, S. (2002). "An unsupervised method for general named entity recognition and automated concept discovery". A: *Proceedings of the 1st International Conference on General WordNet, Mysore, India*.

Asahara, M.; Matsumoto, Y. (2003). "Japanese named entity extraction with redundant morphological analysis". A: *Proceedings of Human Language Technology Conference (HLT-NAACL)* (pàg. 8-15).

Bartschi. M. (1984). *Term Dependencies in Information Retrieval Models*. Tesi doctoral, ETH, Zurich.

Belkin, N.; Croft, W. (1987). *Retrieval technologies*. Elsevier Science.

Bikel, D.; Miller, S.; Schwartz, R.; Weischedel, R. (1997). "Nymble: a high-performance learning name-finder". A: *Proceedings of the fifth conference on Applied natural language processing* (pàg. 194–201). Association for Computational Linguistics Morristown, NJ, EUA.

Black, J. i Ranjan, N. (2004). "Automated event extraction from email". A: *Final Report of CS224N/Ling237 Course in Stanford: http://nlp. stanford. edu/courses/cs224n/2004/, Spring*.

Borthwick, A.; Sterling, J.; Agichtein, E.; Grishman, R. (1998). "NYU: Description of the MENE named entity system as used in MUC-7". A: "*Proceedings of the Seventh Message Understanding Conference (MUC-7)* (vol. 6).

Braschler, M.; Schäuble, P. (2000). "Using Corpus-Based Approaches in a System for Multilingual Information Retrieval". A: *Inf. Retr.*, volum 3(3): págs. 273-284. SIN 1386-4564. Consultable a: http://dx.doi.org/10.1023/A:1026525127581.

Bruandet, M. (1985). "Modele Partiel de Connaisances pour un Systeme de Recherche d'Informations". A: *Proceedings of RIAO85* (pàg. 10–114).

Carreras, X.; Marquez, L. i Padró, L. (2003). "Named entity recognition for Catalan using Spanish resources". A: *Proceedings of EACL'03*.

Clarke, C.; Cormak, G.; Kisman, D.; Lynam, T. (2000). "Question Answering by passage selecion". A: *Proceedings of the Text Retrieval Conference (TREC-9)*.

Croft, W. (1985). "Boolean Queries and Term Dependencies in Probabilistic Retrieval Models". *Journal of the American Society for Information Science* (vol. 37, núm. 2, pàg. 71–77).

Cunningham, H. (2000). *Software Architecture for Language Engineering*. Tesi doctoral, University of Sheffield. Consultable a: http://gate.ac.uk/sale/thesis/.

Dillon, M.; Gray, A. (1983). "FASIT: A Fully Automatic Synstatically Based Indexing System". *Journal of the American Society for Information Science* (vol. 34, núm. 2, pàg. 99–108).

Escudero, G.; Marquez, L.; Rigau, G. (2000). "A Comparision between Supervised Learning Algorithms for Word Sense Disambiguation". A: *Proceedings of the 4th Computational Natural Language Workshop (CoNLL-2000)*.

Ferrández, O.; Kozareva, Z.; Montoyo, A.; Munoz, R. (2005). "Nerua: sistema de detección y clasificación de entidades utilizando aprendizaje automático". *Procesamiento del Lenguaje Natural* (vol. 35, núm. 37-44, pàg. 94).

Fluhr, C. (2000). *Multilingual Information Retrieval*. Consultable a: http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html

Frinkel, J.; Grenager, T.; Manning, C. (2005). "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". A: *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics* (pàg. 363–370).

Gaizauskas, R.; Humphreys, K.; Azzam, S.; Wilks, Y. (1997). "Concepticons vs. lexicons: An architecture for multilingual information extraction". A: *Lecture Notes in Computer Science*, volum 1299 (pàg. 28–43).

Grishman, R. (2003). *Information Extraction* (cap. 30, pàg. 545–599). Oxford University Press.

Hermjakob, U. (2006). *Question Answering from Text, Automatic*, volum 10 (pàg. 323–327). Elsevier.

Ittycheriah, A.; Franz, M.; Zhu, W.; Ratnaparkhi, A.; Mammone, R. (2001). "IBM's statistical question answering system". A: *NIST Special Publication SP* (pàg. 229–234).

Kerkouba, D. (1985). "Indexation Automatique et Aspects Structurels du Texts". A: "Proceedings of RIAO85" (pàg. 227–249).

Kraaij, W.; Pohlmann, R. (1996). "Using Linguistic Knowledge in Information Retrieval". Report tècnic OTS-WP-CL-96001, Research Institute for Language and Speech, Utrech University.

Lesk, M. (1986). "Automatic sense disambiguation using machine redeable dictionaries: how to tell a pine cone from an ice cream cone". A: "ACM Special Interest Group for Design of Communication. Proceedings of the 5th annual international conference on Systems documentation" (pàg. 24–26).

Lin, D.; Pantel, P. (2001). "Induction of semantic classes from natural language text". A: "Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining" (pàg. 317–322). Nova York, EUA: ACM New York.

Manning, C.; Roghavan, P.; Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Martí, M. (2004). *Tecnologías del texto y del habla* (vol. 72 de la Col·lecció UB). Edicions Universitat de Barcelona.

McCallum, A.; Li, W. (2003). "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons". A: *Seventh Conference on Natural Language Learning (CoNLL)*.

Miller, G. (1995). "WordNet: A Lexical Database for English". *Communications of the ACM* (vol. 38, pàg. 11, pàg. 39–41).

Mittendorfer, M. i Winiworter, W. (2002). "Exploiting syntactic analysis of queries for information retrieval". A: *Data & Knowledge Engineering*, volum 42(3).

Moldovan, D.; Harabagiu, S.; Pasca, M.; Mihalcea, R.; Girju, R.; Goodrum, R.; Rus, V. (2000). "The structure and performance of an open-domain question answering system". *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (vol. 38, pàg. 563–570).

Nadeau, D.; Sekine, S. (2009). "A survey of named entity recognition and classification". A: *Named Entities: Recognition, Classification and Use* (pàg. 3).

Nadeau, D.; Turney, P. i Matwin, S. (2006). "Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity". A: *Lecture Notes in Computer Science*, volum 4013 (pàg. 266).

Papp, G. (2009). *Vector-Based Unsupervised Word Sense Disambiguation for Larga Number of Contexts* (pàg. 109–115). TSD 2009, LNAI 5729. Berlin, Heidelberg: Springer-Berlag.

Porter, M. (1980). "An algorithm for suffix stripping". A: *Program*, volum 14(3): (pàg. 120–137).

Ratinov, L.; Roth, D. (2009). "Design Challenges and Misconceptions in Named Entity Recognition". A: *Proceedings of the CoNLL 2009*.

Riloff, E. (1995). "Little words can made a big difference for text classification". A: "Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 95)" (pàg. 130–136).

Robertson, S.; Spärk Jones, K. (1976). "Relevance weighting of search terms". A: *Journal of the American Society for Information Science*, (27) (pàg. 129–146).

Salton, G. (ed.) (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. NJ: Englewood Cliffs. Prentice Hall.

Salton, G. (1989). *Automatic Text Processing - The Transformation, Analysis and Retrieval of Information by Computer*. Reading (MA): Addison-Wesley Publishing Company.

Salton, G. i Lesk, M. (1968). "Computer evaluation of indexing and text processing". *Journal of the ACM* (vol. 15, núm. 1, pàg. 8–36).

Salton, G.; McGill, M. (1983). *Introduction to Modern Information Retrieval*. Nova York: McGraw Hill.

Salton, G.; Wong, A., Yang, C. (1975). "A vector space model for automatic indexing". A: *Communications of the ACM* (vol. 18, núm. 11, pàg. 620).

Sekine, S. (1998). "NYU: Description of the Japanese NE System used for MET-2". A: *Proc. of the Seventh Message Understanding Conference (MUC-7)*. Citeseer.

Shinyama, Y.; Sekine, S. (2004). "Named entity discovery using comparable news articles". A: *Proc. the International Conference on Computational Linguistics (COLING)* (pàg. 848–853).

Shukla, P.; Goyal, P.; Kapil, K.; Mukerjee, A. i Raina, A. (2004). "Multilingual Question Answering". A: "Proceedings of the Symposium on Indian Morphology, Phonology & Language Engineering". Indian Institute of Technology.

Smeaton, A. (1986). "Incorporating syntactic information into a document retrieval strategy: an investigation". A: "Proceedings of the 9th annual internationa ACM SIGIR conference on research and development in information retrieval" (pàg. 103–113).

Soergel, D. (1997). "Multilingual thesauri and ontologies in cross-language retrieval". A: *AAAI Symposium on Cross-language Text and Speech Retrieval*.

Spärk Jones, K. (1999). *What is the role of NLP in text retrieval*. Dordrecht: Kluwer.

Turmo, J.; Català, N.; Rodriguez, H. (1998). "TURBIO: A System for Extracting Information from Restricted Domain Texts". A: *IEA/AIE'98), LNAI 1415* (pàg. 708–721).

Vicedo, J.; Rodriguez Hontoria, H.; Penas, A. i Massot, M. (2003). "Los sistemas de Búsqueda de Respuestas desde una perspectiva actual". A: *Procesamiento del lenguaje natural*, (31): pàg. 351–367.

Wang, Y.; Hoffmann, A. (2006). *Bootstrapping Word Sense Disambiguation Using Dynamic Web Knowledge* (pàg. 1150–1154). Lecture Notes in Computer Science. Springer.

Weiss, S. (2005). *Text mining: predictive methods for analyzing unstructured information*. Springer.

Yarowsky, D. (1995). "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods". A: *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (pàg. 189–196). Cambridge, MA.