Universitat Oberta de Catalunya

DIGITHUM
Les humanitats en l'era digital

http://digithum.uoc.edu

Dossier «Recerca acadèmica sobre la Viquipèdia»

Ús de la Viquipèdia per al desenvolupament de recursos lingüístics: el WordNet 3.0 català i castellà

Antoni Oliver
Professor agregat dels estudis d'Arts i Humanitats i director del postgrau de Traducció i tecnologies de la UOC
aoliverg@uoc.edu

Salvador Climent Roca
Professor agregat dels Estudis d'Arts i Humanitats de la UOC
scliment@uoc.edu

Data de presentació: març de 2012
Data d'acceptació: abril de 2012
Data de publicació: maig de 2012

Resum

En aquest article presentem l'estat de la qüestió en l'ús de la Viquipèdia per a tasques relacionades amb el processament del llenguatge natural i tres aplicacions que hem creat per a l'enriquiment d'un recurs lingüístic de gran abast: el WordNet versió 3.0 per al català i castellà. Els investigadors en aquesta àrea fa anys que cerquen vies perquè les aplicacions integrin informació sobre coneixement del món, d'una manera més o menys estructurada, ja que aquest tipus de coneixement ha demostrat ser molt important per a resoldre de manera satisfactòria moltes tasques de processament del llenguatge. La Viquipèdia pot respondre perfectament a aquesta demanda d'informació amb l'avantatge del seu accés lliure i la seva actualització constant.

Paraules clau

Viquipèdia, WordNet, processament del llenguatge natural, recursos lingüístics

Using Wikipedia to develop language resources: WordNet 3.0 in Catalan and Spanish

Abstract

This paper presents a state of the art on the use of Wikipedia for tasks related to Natural Language Processing and three applications we have developed for enriching a wide-coverage language resource: WordNet version 3.0 for Catalan and Spanish. Researchers in this area have sought for years ways for enriching applications with large quantities of world knowledge in a more or less structured way since it has proved to be crucial for many language processing tasks. Wikipedia can provide such information with some important advantages: free access and constant updating.

Keywords

Wikipedia, WordNet, natural language processing, linguistic resources

Introducció

Un dels objectius fonamentals de la intel·ligència artificial (IA) és dotar les màquines de la capacitat d'entendre el llenguatge humà. Mentre que el processament morfològic i sintàctic de les llengües ha assolit nivells satisfactoris, la comprensió i la representació computacionals del significat continuen essent el gran repte per a la realització de tasques intel·ligents.

A més, des de Langacker (1987) la teoria lingüística postula que el processament semàntic no és específicament lingüístic sinó de naturalesa enciclopèdica, és a dir, inseparable del coneixement del món. En aquesta visió les paraules són punts d'accés a repositoris extensos de coneixement relacionat. Molts investigadors en processament del llenguatge natural (PLN) i IA adopten el mateix punt de partida i cerquen mètodes i fonts d'informació que permetin utilitzar coneixement recopilat i organitzat per humans.

En la darrera dècada, la Viquipèdia (VP) s'ha convertit en la gran alternativa com a font de coneixement semàntic per al PLN. La VP és un corpus de coneixement molt gran, fiable i estructurat. En conseqüència, conté molt més coneixement que qualsevol ontologia o base de coneixement creades manualment –p. ex., la més extensa, CYC (Lenat *et al.,* 1990) o la més utilitzada, WordNet (Fellbaum, 1998)– i també totes les utilitats per al PLN que pugui tenir qualsevol corpus –típicament, l'alimentació de mètodes estadístics–. A més, el format XML, senzill i ben estructurat, i la política de lliurament continuat de *dumps* (paquets amb tota la VP en un tall temporal determinat) simplifica enormement el seu processament.

L'estructura de la VP facilita l'explotació automàtica i semiautomàtica del coneixement que conté. La VP consta d'articles (un article per concepte i un concepte per article), els títols en són representatius i les introduccions, descripcions compactes rellevants i les caixes informatives (*infoboxes*) hi empaqueten estructures de coneixement relacionades. Hi ha enllaços de redireccionament entre termes equivalents i d'hiperenllaç entre termes i altres articles, pàgines de desambiguació entre termes homònims o sinònims i una estructura de categories subjacent a tota la VP en què cada categoria i cada article estan assignats a una o més categories. A més, la VP és àmpliament multilingüe i les estructures de les diverses llengües estan connectades. Finalment, els articles de la VP també poden tenir enllaços a URL externes.

Així doncs, per primer cop en PLN hi ha disponible una font de coneixement explotable que suma qualitat i quantitat d'informació.

En aquest article presentarem en primer lloc l'estat de la qüestió de la recerca en aquest camp i en segon lloc una investigació pròpia en què s'utilitzen estratègies multilingües per a l'extracció de coneixement semàntic monolingüe.

1. Estat de la qüestió

Thomas i Amit (2007) han demostrat que la qualitat de la informació continguda pels articles de la VP es pot inferir del seu historial d'edició, ja que hi ha una correlació clara entre l'estabilització de la producció de canvis i la qualitat, la qual, per cert, és equiparable a la de l'*Enciclopaedia Britannica* (cf. Giles, 2005) però deu cops més voluminosa –l'any 2009.

Exposarem a continuació els principals usos que es dóna a la VP en recerca en PLN a partir dels estats de la qüestió proporcionats per Medelyan *et al*. (2009), Gabrilovich *et al.* (2009) i Vivaldi *et al.* (2010).

1.1. Desambiguació lèxica, similitud semàntica i classificació automàtica de documents

Raonar sobre la similitud de dues paraules, frases o documents és una tasca rutinària per als humans, però un repte lluny de l'abast de les màquines. El mètode més habitual per a establir la similitud entre dos documents és a partir d'anàlisis estadístiques del grau de coincidència dels mots que contenen, ponderades amb l'ocurrència d'aquests mots en corpus textuals (*Latent Semantic Analysis*, LSA; Deerwester *et al*. 1990). Gabrilovich *et al.* (2007) han desenvolupat una tècnica que calcula el grau d'associació entre els mots d'un document i entrades de la VP, de manera que la comparació d'aquesta representació amb la d'un altre document en dóna una estimació del grau de similitud que ha demostrat ser molt més acurada que el LSA.

Aquest càlcul de similitud pot ser útil per a moltes tasques, com per exemple el descobriment del plagi o la classificació automàtica de documents, i aquesta darrera tasca serveix al seu torn per a l'encaminament de notícies o correus electrònics o la identificació de correu brossa.

Un cas límit de raonament sobre similitud de cadenes de text és la desambiguació del significat lèxic, és a dir, decidir si dos mots iguals en forma tenen o no el mateix significat. Aquesta és una de les tasques més necessàries en totes les aplicacions de recuperació, extracció i tractament de la informació i també de les més difícils, ja que el llenguatge és essencialment ambigu: els humans desambigüem contínuament entre els sentits possibles de cada mot a partir del context en què ocorren. La base de coneixement més utilitzada en PLN com a inventari de sentits de paraules organitzats en conceptes relacionats és WordNet (Fellbaum, 1998), el qual presenta els problemes principals de l'excessiva proliferació de significats per mot i del biaix cap al lexicó comú –i, per tant, baix contingut de noms propis i termes especialitzats.

Diversos autors, com Medelyan *et al*. (2008), fan servir tècniques com les descrites suara per al càlcul de la similitud de documents i les apliquen a la comparació entre, d'una banda, els conceptes candidats a ser el significat del mot analitzat (conceptes

representats per la seva entrada de VP) i, d'altra banda, els mots que en constitueixen el context en l'oració on apareix.

Quant al lèxic propi i especialitzat, actualment la VP n'és el repositori més ampli que es coneix. Diverses recerques enllacen mots dels textos a articles de la VP i així realitzen càlculs de desambiguació i categorització, com per exemple, en un determinat context, decidir que *Chicago* és el grup musical i no la ciutat ni l'espectacle de Broadway.

Una altra tasca lingüística necessària en totes les aplicacions intel·ligents és saber quan diferents mots del text (per exemple *el rei* i *Joan Carles I*) refereixen al mateix concepte. Per a resoldre el problema, a més d'aplicar mesures de similitud com les abans descrites, autors com Ponzetto *et al.* (2006) exploten l'aparició dels mots al primer paràgraf de la VP, els seus enllaços o les llistes de categories VP comunes per a decidir sobre la identitat de significat.

La classificació automàtica de documents és una tasca que vol organitzar automàticament col·leccions de documents en categories preestablertes, habitualment a partir dels mots que contenen. Aquesta aproximació simple funciona malament quan documents diferents parlen dels mateixos temes amb diferents paraules. Gabrilovich *et al.* (2006) i altres treballs optimitzen la tasca augmentant les llistes de paraules dels documents amb conceptes relacionats extrets de la VP. Banerjee (2007) fa notar que aquesta estratègia permet que els mètodes siguin més adaptables, ja que doten la informació d'una base més estable independentment de l'evolució constant de la terminologia especialitzada.

1.2. Recuperació d'informació i obtenció de respostes

La recuperació d'informació i l'obtenció de respostes a partir de col·leccions de documents és una tasca que avui en dia fem constantment, en la versió que utilitza els cercadors populars com a eina de pregunta i internet com a col·lecció de documents. La VP s'utilitza per a millorar aquesta tasca en diferents sentits.

En primer lloc, amb la VP es poden expandir automàticament les preguntes de cerca amb sinònims, formes alternatives i conceptes relacionats per a obtenir documents més rellevants. Aquestes operacions, que ja s'han fet abans amb WordNet, es veuen optimitzades amb la VP (p. ex., Gregorowicz *et al.*, 2006) a causa de la major extensió de la VP i del seu caràcter dinàmic, ja que la VP creix constantment en termes i cobertura de tecnologies emergents, mentre que els altres recursos són estàtics o evolucionen molt lentament.

L'obtenció de respostes (*Question Answering*) és una variant de la recuperació d'informació en la qual s'espera obtenir com a resultat de la pregunta no pas una col·lecció de documents sinó una frase o un fragment rellevant de document que li doni resposta específica. Diversos autors, com Kaisser (2008), incorporen a l'habitual cerca a internet informació estructurada de VP que ajuda a establir un factor clau: el tipus d'entitat que se cerca com a resposta.

Amb relació a la cerca intel·ligent a la web, Wu *et al.* (2007) i altres han proposat revitalitzar el projecte del web semàntic (Berners-Lee *et al*. 2001), que fins ara s'ha demostrat inviable atesa la magnitud de la tasca de marcar semànticament tot el web. Aquests autors proposen etiquetar semànticament de manera clara la pròpia estructura d'hiperenllaços de la VP de manera que es converteixi en una estructura de coneixement «llegible» per les màquines i constitueixi així el nucli d'un futur web semàntic real.

1.3. Expansió i millora de bases de coneixement semàntic

L'estructura de la VP facilita el manteniment o l'expansió d'estructures de coneixement, com tesaurus o ontologies, imprescindibles per a moltes tasques i aplicacions. L'explotació d'enllaços jeràrquics i de redirecció de la VP permeten establir sinònims i relacions d'inclusió entre conceptes i descobrir conceptes nous no coberts per altres ontologies. Les *infoboxes* i l'estructura de categories permeten convertir directament la VP en bases de dades relacionals. A més, el primer paràgraf dels articles permet establir glosses o descripcions dels conceptes. Els enllaços interlingüístics fan possible la creació de bases de coneixement equivalents en noves llengües i la comparació de l'estructura de la VP amb la d'altres ontologies permet l'ampliació d'ambdues.

L'extracció d'informació és una tasca que cerca constituir bases de coneixement estructurades a partir de text no estructurat. L'explotació de la naturalesa enciclopèdica de la VP i del seu estil uniforme de redacció optimitza aquesta tasca.

De manera més ambiciosa, l'explotació de l'estructura d'enllaços i categories de la VP permet avançar cap a una representació explícita de fets i regles augmentada amb sistemes d'herència que permeti capturar l'estructura semàntica global del llenguatge. Aquesta estructura, que es pot extreure de la VP, pot arribar a ser una alternativa més econòmica i amb més cobertura que les bases de coneixement desenvolupades manualment, com WordNet i CYC. Les tres principals aproximacions que s'investiguen ara mateix són: l'extracció i etiquetatge de les relacions categorials de la VP, l'extracció i organització de la informació continguda en les *infoboxes* i l'enriquiment dels recursos manuals amb informació extreta de la VP.

El principal projecte en aquest sentit és YAGO (Suchanek *et al*., 2007), una taxonomia de grans dimensions creada enriquint l'estructura de conceptes de WordNet amb categories i articles de la VP. A més, aquests autors enriqueixen WordNet amb relacions semàntiques que no hi eren, però sí a la VP, com «nascut_el_any» o «nascut_a_lloc». Amb aquest procés els autors asseguren haver creat una base de coneixement de 20 milions de fets atribuïts a 2 milions d'entitats, dotada, a més, d'una descripció lògica prou expressiva per a fer inferències a partir d'aquests fets. Un altre projecte en el mateix sentit és DBpedia (Auer *et al*., 2007) que ha transformat l'estructura de les *infoboxes* en un conjunt gegant (103 milions) de proposicions lògiques.

1.4. Aplicacions educatives

Sawaki *et al*. (2007) han desenvolupat una variant de la cerca de respostes que pot tenir aplicació docent: extreuen d'articles biogràfics de la VP preguntes, respostes i pistes amb les quals generen automàticament enigmes del tipus «qui és aquest personatge?». Una línia de treball interessant respecte a això és l'ordenació de les pistes per dificultat, de manera que l'enigma es pugui plantejar de manera que n'augmenti el grau de dificultat.

A casa nostra, Moré (2009) explota l'estructura jeràrquica de categories de la VP per a generar camps semàntics de conceptes (que poden ser multilingües) aplicables a l'elaboració de materials didàctics i altres documents acadèmics. També, Moré *et al*. (2010) utilitzen la VP com a font d'informació de suport per a docents en línia, utilitzant tècniques d'obtenció de respostes que fan servir missatges dels alumnes com a preguntes, i VP i altres fonts com a col·lecció de documents on es poden trobar les respostes.

1.5. Multilingüisme

Pel seu caràcter multilingüe i la connexió interlingüística de les entrades, la VP és alhora un diccionari plurilingüe i un corpus alineat, per la qual cosa és susceptible de ser utilitzat en qualsevol tasca relacionada amb la traducció automàtica o la recuperació d'informació interlingüística (recuperar dels cercadors informació expressada en una llengua diferent a la llengua en què es fa la consulta) mitjançant la creació i expansió multilingüe de preguntes.

Una altra explotació interessant és la d'Adafre *et al*. (2006), que, treballant en anglès i holandès, comparen traduccions automàtiques d'articles de la VP amb el corresponent article escrit a mà i a partir d'això generen lexicons bilingües, els quals utilitzen per a identificar parells d'oracions en llengües diferents que tenen un significat igual o similar. Al seu torn, Erdmann *et al*. (2008) fan explotacions sofisticades de l'estructura de la VP, més enllà dels simples enllaços interlingüístics directes, per a crear diccionaris bilingües.

A continuació presentarem un treball propi que s'insereix en aquesta darrera línia, la menys explotada actualment, d'explotació del multilingüisme de la VP per a la creació de bases de coneixement.

2. WordNet

Com hem vist a l'apartat anterior, WordNet (Fellbaum, 1998) és una base de dades de coneixement lèxic de l'anglès. En aquesta base de dades les paraules de les categories obertes (substantius, verbs, adjectius i adverbis) s'organitzen en conjunts de sinònims que reben el nom de *synsets*. Cada *synset* representa un concepte lexicalitzat en anglès, i es connecta amb els altres *synsets* mitjançant relacions semàntiques. Les principals relacions que ofereix el WordNet són la *hiponímia* o relació d'especificitat entre un mot (l'hipònim) i un altre de significat més genèric (hiperònim), l'*antonímia* o relació entre mots que tenen un significat directament oposat, la *meronímia* o la relació entre una part i un tot, i la *troponímia* o implicació lèxica, una relació que es dóna entre verbs i que es pot considerar en certa manera equivalent a la relació d'hiponímia per als substantius.

Per exemple, el *synset* del WordNet 3.0 de l'anglès que s'identifica mitjançant un *offset* i una categoria gramatical 02958343-n té diverses *variants*: *car*, *auto*, *automobile*, *machine* i *motorcar*. Cada *synset* té assignada una glossa o definició i molt sovint exemples d'ús. Per al *synset* d'exemple és «a motor vehicle with four wheels; usually propelled by an internal combustion engine»; i l'exemple és «he needs a car to get to work». Aquest *synset* té 31 hipònims, com per exemple 02701002-n (*ambulance*) o 03594945-n (*jeep, landrover*), entre d'altres. També té un hiperònim, el 03791235-n (*motor vehicle, automobile vehicle*). D'entre els meròmins registrats a WordNet podem posar com a exemple el 02685365-n (*airbag*).

Per donar un exemple de troponímia hem de considerar algun *synset* verbal, com per exemple, 01926311-v («run; move fast by using one's feet. with one foot off the ground at any given time»). La relació de troponímina es pot concretar mitjançant una sèrie d'hipònims (12 segons WordNet), per exemple 01928579-v («sprint; run very fast, usually for a short distance») i hiperònims (1 segons WordNet) 02055649-v («travel rapidly, speed, hurry, zip; move very fast»).

La relació de sinonímia es dóna entre totes les formes lèxiques (*variants*) d'un determinat *synset* (per exemple, *car*, *auto*, *automobile*, *machine* i *motorcar* son sinònims, ja que totes aquestes paraules són *variants* del *synset* 02958343-n).

WordNet ha esdevingut un recurs estàndard per a tot tipus de recerques i aplicacions semàntiques en l'àrea del PLN. El WordNet anglès és lliure i es pot baixar de la seva pàgina web de la Universitat de Princeton (http://wordnet.princeton.edu). En la resta d'aquest article a aquest WordNet l'anomenarem PWN (*Princeton WordNet*). La versió actual és la 3.0, publicada el desembre de 2006. A la taula 1 podem observar una comparativa del nombre de *synsets* per a les versions 1.5, 1.6 i 3.0 del PWN.

	1.5	1.6	3.0
Total	76.705	99.642	118.695
Substantius	51.253	66.025	83.073
Verbs	8.847	12.127	13.845
Adjectius	13.460	17.915	18.156
Adverbis	3.145	3.375	3.621

Taula 1: Nombre de synsets per a diferents versions del PWN

Com podem observar, el nombre de *synsets* desenvolupats augmenta amb les noves versions. Aquest fet obliga a actualitzar també els WordNets per a altres llengües, amb l'objectiu que els WordNets siguin comparables.

No tots els WordNets que es construeixen s'editen amb una llicència lliure. A Bond (2012) podem trobar els WordNets existents per a les diferents llengües amb la llicència associada. Les llengües que disposen de WordNets lliures són: anglès, finès, rus, tailandès, danès, japonès, català, gaèlic, hindi, francès, malai, indonesi, castellà, àrab i hebreu.

2.1. Estratègies de construcció de WordNets

En aquest apartat repassarem algunes de les estratègies que s'han fet servir per a la construcció de WordNets per a diverses llengües (sense incloure l'anglès, ja que el WordNet anglès es considera l'original). Vossen (1998) distingeix dues aproximacions generals per a la construcció de WordNets:

- **Estratègia de combinació** (*merge model*): es genera una ontologia pròpia per a cada llengua i posteriorment es generen les relacions interlingüístiques entre el WordNet generat i el PWN.
- **Estratègia d'expansió** (*expand model*): es tradueixen les *variants* associades als *synsets* del PWN fent servir diverses estratègies. En aquest cas no és necessari establir relacions interlingüístiques perquè el WordNet generat i el PWN són paral·lels.

Cada una d'aquestes estratègies presenta avantatges i inconvenients (Vossen, 1996). L'estratègia d'expansió és tècnicament més senzilla i garanteix un grau més alt de compatibilitat entre els WordNets de les diferents llengües. Però els WordNets desenvolupats d'aquesta manera estan molt influenciats pel PWN i contindran tots els seus errors i deficiències estructurals. L'estratègia de combinació és més complexa però permet un major aprofitament més directe de les ontologies i tesaurus disponibles.

2.2. Els WordNets del català i el castellà

Els primers WordNets per al català (Benítez *et al.*, 1998) i el castellà (Atserias *et al.*, 1997) es van construir a partir del PWN 1.5 seguint una estratègia d'expansió, ja que va consistir en la traducció de les *variants* corresponents al *synsets* del PWN. Per a fer aquesta traducció es va fer servir principalment una tècnica basada en diccionaris bilingües. Els diccionaris proporcionen la relació entre les paraules angleses i les catalanes (o castellanes), però no entre els synsets i les paraules catalanes. Per a poder establir nivells de confiança en l'assignació de variants catalanes a *synsets* van dividir en diversos grups les relacions: (i) entre paraules angleses i *synsets* en dos grups, monosèmiques i polisèmiques, i (ii) entre paraules angleses i catalanes depenent del nombre de traduccions que tenen. Això permet fer una relació directa entre *synsets* i paraules catalanes (i s'obté d'aquesta manera les *variants* en català) i assignar a cada relació un nivell de confiança.

Posteriorment es van generar les versions 1.6 dels WordNets català i castellà mitjançant un *mapping* que relaciona els *synsets* de la versió 1.5 amb els de la versió 1.6.

La gran diferència entre aquests dos WordNets ha estat la llicència amb la qual s'ha distribuït: el WordNet català s'ha distribuït íntegrament amb una llicència lliure (GNU-GPL), mentre que el castellà s'ha distribuït amb una llicència propietària, tot i que un fragment reduït s'ha distribuït lliurement juntament amb l'analitzador Freeling i que la versió íntegra estava disponible lliurement per a recerca.

3. Ús de la Viquipèdia per a la construcció de WordNets

En aquest apartat presentem tres aplicacions basades en la Viquipèdia que hem desenvolupat per a l'enriquiment del WordNet 3.0 del català i castellà amb noves variants o conceptes. En primer lloc explotem les dades d'una base de coneixement lligada a la Viquipèdia, Babelnet; en segon lloc s'utilitza la Viquipèdia com a diccionari bilingüe i en tercer lloc com a font de noms propis invariables.

3.1. El projecte Babelnet

L'objectiu del projecte BabelNet (Navigli *et al.*, 2010) és crear una xarxa semàntica de grans dimensions que relaciona el coneixement lexicogràfic de WordNet i el coneixement enciclopèdic de la Viquipèdia. A la figura 1 podem veure com Babelnet relaciona el *synset* 02958343-n, que té les *variants motorcar, automobile i car* amb l'entrada de la Viquipèdia anglesa que porta per títol *Car*.

Figura 1. Relació entre el synset 02958343n i la seva corresponent entrada de la Viquipèdia

5558	Motorcar	motorcar%1:06:00::	02958343n
11802	Automobile	automobile%1:06:00::	02958343n
45193	Car	car%1:06:00::	02958343n

Partint d'aquest treball, gràcies al qual disposem de la relació entre un *synset* i una entrada de la Viquipèdia anglesa *weng*, podem fer sevir els enllaços interlingüístics per a establir la mateixa relació per a la resta de llengües que disposin de l'entrada equivalent *wcat, wspa,* etc.

A la figura 2 podem veure els enllaços interlingüístics de l'entrada *Car*, extrets d'un *dump* de la Viquipèdia en XML. Amb aquesta informació podem relacionar el *synset* 02958343n amb la variant catalana *automòbil*.

Figura 2: Fragment dels enllaços interlingüístics corresponents a l'entrada

[[ast:Automóvil]]
[[gn:Mba'yruguata]]
[[az:Avtomobil]]
[[bn:গাড়ী]]
[[zh-min-nan:Chū-tōng-chhia]]
[[be:Аўтамабіль]]
[[be-x-old:Аўтамабіль]]
[[bs:Automobil]]
[[br:Karr-tan]]
[[bg:Автомобил]]
[[ca:Automòbil]]
[[cs:Automobil]]
[[co:Vittura]]
[[cy:Car]]
[[da:Bil]]
[[pdc:Maschien]]
[[de:Automobil]]
[[nv:Chidí]]
[[et:Auto]]
[[el:Αυτοκίνητο]]
[[es:Automóvil]]

A continuació podem buscar també altres variants de l'entrada original. La Viquipèdia conté una sèrie de pàgines de redirecció que readrecen certes entrades a una entrada principal, a la qual són equivalents. Com podem veure a la figura 3, que mostra l'estructura en XML de l'entrada *Cotxe*, aquesta entrada es redirigeix cap a la d'*Automòbil*. És a dir, si cerquem *Cotxe* a la Viquipèdia catalana, el que fa el sistema és mostrar la informació corresponent a l'entrada *Automòbil*. Aquest sistema de redireccions ens permet establir, per exemple, que *cotxe* és una *variant* catalana vàlida per al *synset* 02958343n.

Figura 3. Informació referent a l'entrada Cotxe de la Viquipèdia

```
<page>
  <title>Cotxe</title>
  <id>19827</id>
  <redirect />
  <revision>
    <id>124331</id>
    <timestamp>2004-10-14T07:13:01Z</timestamp>
    <contributor>
      <username>Arnadí</username>
      <id>281</id>
    </contributor>
    <comment>#REDIRECT [[Automòbil]]</comment>
    <text xml:space="preserve">#REDIRECT [[Automòbil]]</text>
  </revision>
</page>
```

A la pràctica, però, és difícil aprofitar aquesta característica, ja que les pàgines de redirecció tenen un ús real més ampli. Per exemple, d'*Automòbil* en trobaríem els sinònims següents: *Automoció, Cotxe, Cotxes, Elements aerodinàmics en l'automòbil.* Aquest exemple ens mostra que aquesta informació no pot ser aprofitada directament de manera automàtica, però sí que pot resultar d'utilitat sempre i quan es faci una posterior revisió manual de les propostes.

En alguns casos pot passar que la llengua d'interès no disposi de la corresponent entrada de la Viquipèdia. Per a poder obtenir *variants* en aquesta llengua, Navigli *et al.* (2010) fan servir el sistema de traducció automàtica Google Translate per a traduir una sèrie d'oracions en anglès que continguin *variants* dels *synsets*. Aquestes oracions les extreuen de les dues fonts següents:

- El corpus Semcor (Miller *et al*., 1993), un corpus de l'anglès etiquetat semànticament; és a dir, que té els substantius, verbs, adjectius i adverbis (si no tots, almenys la majoria) amb una etiqueta que indica el seu sentit. Les etiquetes d'aquest corpus són els *synsets* del WordNet. A la Figura 4 podem observar un exemple de frase d'aquest corpus amb les seves etiquetes (el format real del corpus és diferent). El corpus Semcor es pot baixar de http://www.cse.unt.edu/~rada/downloads.html.

Figura 4. Exemple d'oració del corpus Semcor (amb un format adaptat per a una millor comprensió)

Then/00117620r he noticed/0215408v that the dry/02551380a wood/15098161n of the wheels/0457499n had swollen/00256507v.

- Oracions de la Viquipèdia que contenen enllaços a la pàgina *weng*. Els textos de la Viquipèdia contenen molts hiperenllaços a pàgines de la mateixa Viquipèdia. A la figura 5 podem veure un fragment de l'entrada d'*Automòbil* amb els hiperenllaços marcats en color blau. En certa manera aquests hiperenllaços es poden considerar també un etiquetatge semàntic, ja que per a paraules ambigües enllacen a la pàgina corresponent al sentit correcte. Com a exemple, podem considerar a la figura 5 la paraula *bateria*, que és ambigua. La Viquipèdia ens dóna informació sobre les ambigüitats de les paraules en les *pàgines de desambigüació*; a la figura 6 podem observar la informació de desambigüació de *bateria*. Atès que l'enllaç de la figura 5 apunta a la pàgina *Bateria elèctrica* podem saber que el significat d'aquesta paraula en el text és precisament aquest.

Figura 5. Fragment de l'entrada Automòbil amb els hiperenllaços marcats en blau

L'**automòbil** (comunament **cotxe** o **votura** a la Catalunya Nord) és usualment un vehicle de quatre rodes destinat al transport de persones, amb capacitat entre dos i vuit seients. Es desplaça gràcies a un motor d'explosió a base d'una mescla de gasolina, gasoil i aire. En alguns països el combustible es fabrica a partir de determinades plantes en forma d'alcohol etílic. Recentment s'han començat a produir automòbils que funcionen amb motor elèctric, si bé l'autonomia d'aquests vehicles és encara limitada a causa del pes de les bateries. Les rodes davanteres dels automòbils poden moure's cap a ambdós costats per a fer girs i prendre les corbes.

Figura 6. Informació de desambiguació de la paraula bateria

Bateria té els significats següents.

- *Electricitat:* Bateria elèctrica, conjunt d'acumuladors connectats en sèrie o en paral·lel
- *Exèrcit:* Bateria (unitat militar), agrupació tàctica i de tir elemental composta per un conjunt d'artillers i de canons
- *Música*: Bateria (instrument musical), conjunt d'instruments de percussió que és tocat per un sol instrumentista
- Eines: Conjunt d'estris de cuina. Vegeu la Categoria: Estris de cuina

Un cop feta la traducció automàtica d'aquestes oracions, identifiquem la traducció més freqüent i la incloem com a *variant* corresponent per a la llengua d'interès.

3.2. Ús de la VP com a diccionari bilingüe

A partir de la Viquipèdia podem generar diccionaris bilingües de tipus enciclopèdic; és a dir, a més d'algunes paraules de la llengua general que tenen interès enciclopèdic, hi trobarem noms propis de persones, llocs geogràfics, etc. Per a confeccionar aquests diccionaris simplement hem de seguir els enllaços interlingüístics. Per a tenir una idea de magnitud, a partir de la Viquipèdia anglesa podem confeccionar un diccionari bilingüe anglès-català de 233.130 entrades i anglès-castellà de 481.105 entrades.

Aquest tipus de diccionaris poden servir per a fer una assignació directa de *variant* catalana o castellana a partir de les variants angleses assignades a un determinat *synset*. Aquesta assignació només es pot determinar amb certa seguretat en els casos que una determinada variant estigui assignada a un únic *synset*, és a dir, que la paraula sigui monosèmica. A la taula 2 podem observar una estadística sobre la monosèmia i polisèmia de les *variants* del WN3.0 de l'anglès. Com podem observar, la majoria de les variants presents al WN3.0 anglès són monosèmiques, concretament el 82,32% de les variants.

Nombre de sentits	Variants	%
1	123.228	82,32
2	15.577	10,41
3	5.027	3,36
4	2.199	1,47
5 o més	3.659	2,44
Total variants	149.691	100

Taula 2: Monosèmia i polisèmia de les variants del WN 3.0 anglès

D'aquestes *variants* monosèmiques un cert percentatge tindran correspondència en els enllaços interlingüístics (dels quals hem pogut determinar el diccionari bilingüe) i podrem assignar directament una variant catalana o castellana (que convindria revisar manualment). Concretament, hem generat amb aquest mètode 3.719 variants per al català i 5.260 per al castellà.

3.3. Detecció de noms propis invariables

WordNet conté una gran quantitat de *variants* corresponents a noms propis. Si ens fixem en les *variants* escrites amb alguna lletra en majúscula observem que la versió 3.0 de l'anglès en conté més de 40.000. Moltes d'aquestes variants corresponen a noms de persones o llocs geogràfics que s'escriuen exactament igual en anglès, català i castellà, però que no estan recollits com a entrades de la Viquipèdia catalana o castellana i per tant no tenen enllaços interlingüístics cap a aquestes llengües.

Podem fer servir els enllaços interlingüístics cap a altres llengües per a verificar si, a més d'escriure's així en anglès, la *variant* en qüestió també s'hi escriu en un nombre de llengües que determinem, per exemple 5. En els primers experiments en què hem fet

servir aquesta estratègia, i que hem dut a terme únicament amb unitats multiparaula, s'han obtingut 2.858 *variants* per al català i 2.928 per al castellà amb precisions superiors al 87%.

4. Conclusions

En aquest article hem presentat l'estat de la qüestió sobre l'ús de la Viquipèdia en tasques relacionades amb el processament del llenguatge natural i més concretament per a la creació de recursos lingüístics. Relacionat amb aquest darrer aspecte hem presentat tres desenvolupaments en què fem ús de la Viquipèdia per a la generació automàtica de les bases de coneixement lèxic WordNet 3.0 per al català i castellà, recurs que s'ha convertit en estàndard per al processament semàntic. Les propostes presentades tenen l'avantatge que poden ser aplicades a qualsevol llengua que disposi de versió de la Viquipèdia i, tot i que naturalment no permeten la creació de WordNets complets, poden ser molt útils si es combinen amb altres tècniques, manuals o automàtiques.

Referències bibliogràfiques

ADAFRE, S. F.; DE RIJKE, M. (2006). «Finding Similar Sentences across Multiple Languages in Wikipedia». A: *Proceedings of EACL 2006 Workshop on New Text, Wikis and Blogs and Other Dynamic Text Sources*. Pàg. 62-69. Trento, Itàlia.

ATSERIAS, J.; CLIMENT, S.; FARRERES, X.; RIGAU, G.; RODRÍGUEZ, H. (1997). «Combining multiple methods for the automatic construction of multilingual wordnets». A: *Proceedings of RANLP'97*. Pàg. 143-149. Tsigov Chark, Bulgària.

AUER, S.; BIZER, C.; LEHMAN, J.; KOBILAROV, G.; CYGANIAK, R.; IVES, Z. (2007). «DBpedia: A Nucleus for a Web of Open Data». Aberer [*et al*.] (editors): *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea. Lecture Notes in Computer Science 4825 Springer 2007*.

BANERJEE S. (2007). «Boosting inductive transfer for text classification using Wikipedia». A: *Proceedings of the 6th International Conference on Machine Learning and Applications* (ICMLA), pàg. 148-153. Cincinnati, Ohio, USA.

BENÍTEZ, L; CERVELL, S.; ESCUDERO, G.; LÓPEZ, M.; RIGAU, G.; TAULÉ, M. (1998). «Methods and tools for building the catalan wordnet». A: *Proceedings of the ELRA Workshop on Language Resources for European Minority Languages, First International Conference on Language Resources & Evaluation*. Granada, Espanya.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O.(2001). «The semantic web». *Scientific American*. Vol. 284, núm. 5, pàg. 34-43. http://dx.doi.org/10.1038/scientificamerican0501-34

BOND, F.; KYONGHEE, P. (2012). «A Survey of WordNets and their Licenses». A: *Proceedings of the 6th International Global WordNet Conference*. Matsue, Japó. Pàg. 64-71.

DEERWESTER, S.; DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K.; HARSHMAN, R. (1990). «Indexing By Latent Semantic Analysis». *Journal of the American Society For Information Science*. Núm. 41, pàg. 391-407. http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

ERDMANN, M.; NAKAYAMA, K.; HARA, T.; NISHIO, S.(2008). «An Approach for Extracting Bilingual Terminology from Wikipedia». A: *Proceedings of International Conference on Database Systems for Advanced Applications (DASFAA)*. Nova Delhi, la Índia. http://dx.doi.org/10.1007/978-3-540-78568-2_28

FELLBAUM, C. (1998). «WordNet: An Electronic Lexical Database and some of its Applications». MIT Press.

GABRILOVICH, E.; MARKOVITCH, S. (2006). «Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge». A: *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*. Boston. Pàg. 1301-1306.

GABRILOVICH, E.; MARKOVITCH S. (2007). «Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis». A: *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI'07*. Hyderabad, Índia. Pàg. 1606-1611.

GABRILOVICH, E.; MARKOVITCH, S. (2009) «Wikipedia-based Semantic Interpretation for Natural Language Processing». *Journal of Artificial Intelligence Research*. Núm. 34, pàg. 443-498.

GREGOROWICZ, A.; KRAMER, M. A. (2006). «Mining a Large-Scale Term-Concept Network from Wikipedia». Informe tècnic 06-1028, Mitre.

GILES, J. (2005). «Internet encyclopaedias go head to head». *Nature*. Núm. 438, pàg. 900-901. http://dx.doi.org/10.1038/438900a

KAISSER, M. (2008). «The QuALiM Question Answering Demo: Supplementing Answers with Paragraphs drawn from Wikipedia». A: *Proceedings of ACL (Demo Papers)*. Pàg. 32-35.

LANGACKER, R. (1987). *Foundations of Cognitive Grammar, Volume I.* Stanford, Califòrnia: Stanford University Press.

LENAT, D.; GUHA, R.V. (1990). *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley.

MEDELYAN, O.; WITTEN, I.H.; MILNE, D. (2008). «Topic Indexing with Wikipedia». *Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy.* Chigago: AAAI Press.

MEDELYAN, O.; MILNE, D.; LEGG, C.; WITTEN, I.H. (2009). «Mining Meaning from Wikipedia». *International Journal of Human-Computer Studies.* Núm. 67, pàg. 716-754. http://dx.doi.org/10.1016/j.ijhcs.2009.05.004

MILLER, G.A. [et al.] (1993). «A semantic concordance». A: *Proceedings of the Workshop on Human Language Technology*. pp. 303–308. Stroudsburg, Pennsylvania, USA. http://dx.doi.org/10.3115/1075671.1075742

MORÉ, J. (2009). «Creació automàtica de diccionaris multilingües especialitzats en noves àrees temàtiques» [article en línia]. *Digithum*. Núm. 11. UOC. [Data de consulta: 27/02/2012]. <http://www.uoc.edu/ojs/index.php/digithum/article/view/n11_more/n11_more>

MORÉ, J.; CLIMENT, S.; COLL-FLORIT, M.; RIVERA, J. (2010). «A Question-Answering Environment for eLearning Tutors». A: *International Journal of the Computer, the Internet and Managment (IJCIM)*. Vol. 18, núm. SP1, pàg. 3.1-3.6. ISSN 0858-7027. <http://www.ijcim.th.org/v18nSP1.htm>

NAVIGLI, R.; PONZETTO, S. P. (2010). «BabelNet: building a very large multilingual semantic network». A: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL'10. Pàg. 216-225. Stroudsburg, PA, EUA: Association for Computational Linguistics. [Data de consulta: 11/3/2011]. <http://portal.acm.org/citation.cfm?id=1858681.1858704>

PONZETTO, S. P.; STRUBE, M. (2006). «Exploiting semantic role labelling, WordNet and Wikipedia for coreference resolution». A: *Proceedings of HLT-NAACL'06*. Pàg. 192-199.

SAWAKI, M.; MINAMI, M. Y.; HIGASHINAKA, R.; DOHSAKA, K.; YAMADA, T.; MATSUBAYASHI, T.; ISOZAKI, H.; MAEDA, E. (2007). «Quizmaster Mushrooms: 'Who is this' Quiz Dialogue System». International Conference on Multimodal Interaction ICMI (Demostració). Nagoya, Japó.

SUCHANEK, F. M.; KASNECI, G.; WEIKUM, G. (2007). «Yago - A Core of Semantic Knowledge». A: *Proceedings of the 16th international World Wide Web conference (WWW 2007)*. Nova York: ACM Press.

THOMAS, C. S.; AMIT, P.(2007). «Semantic Convergence of Wikipedia ARrticles». A: *Proceedings of the International Conference on Web Intelligence* (IEEE/WIC/ACM WI'07). Hong Kong.

VIVALDI, J.; RODRIGUEZ, H. (2010). «Finding Domain Terms using Wikipedia». A: *Proceedings of the 7th LREC International Conference*. Malta. Pàg. 386-393.

VOSSEN, P. (1996). «Right or Wrong. Combining lexical resources in the EuroWordNet project». A: *Proceedings of Euralex-96*. Göteborg. Pàg. 715-728.

VOSSEN, P. (1998). «Introduction to Eurowordnet». *Computers and the Humanities.* Vol. 32, núm. 2, pàg. 73-89. http://dx.doi.org/10.1023/A:1001175424222

WU, F.; WELD, D. (2007). «Autonomously semantifying Wikipedia». A: *Proceedings of the 16th ACM Conference on Information and Knowledge Management, CIKM'07*. Lisboa: Portugal. Pàg. 41-50.

CITACIÓ RECOMANADA

OLIVER, Antoni; CLIMENT, Salvador (2012). «Ús de la Viquipèdia per al desenvolupament de recursos lingüístics: el WordNet 3.0 català i castellà». A: Eduard Aibar i Mayo FUSTER (coords.). «Recerca acadèmica sobre la Viquipèdia» [dossier en línia]. *Digithum,* núm. 14, pàg. 15-24. UOC. [Data de consulta: dd/mm/aa]. <http://digithum.uoc.edu/ojs/index.php/digithum/article/view/n14-oliver-climent/n14-oliver-climent-cat> http://dx.doi.org/10.7238/d.v0i14.1474 ISSN 1575-2275

Antoni Oliver
Professor agregat dels estudis d'Arts i Humanitats i director del postgrau
de Traducció i tecnologies de la UOC
aoliverg@uoc.edu

Antoni Oliver és professor agregat dels estudis d'Arts i Humanitats de la Universitat Oberta de Catalunya i director del postgrau en Traducció i tecnologies. La seva àrea de recerca és la lingüística computacional, especialment en àmbits relacionats amb la traducció automàtica.

Estudis d'Arts i Humanitats
Universitat Oberta de Catalunya
Avinguda Tibidabo 39-43
08035 Barcelona

Salvador Climent Roca
Professor agregat dels Estudis d'Arts i Humanitats de la UOC
scliment@uoc.edu

Salvador Climent és professor agregat dels Estudis d'Arts i Humanitats de la Universitat Oberta de Catalunya, on dirigeix el grau de Llengua i Literatura Catalanes i imparteix i coordina assignatures de lingüística general i cognitiva. Fa recerca en lingüística computacional i lingüística cognitiva.

Estudis d'Arts i Humanitats
Universitat Oberta de Catalunya
Avinguda Tibidabo 39-43
08035 Barcelona