

SENSEVAL, una aproximació computacional al significat



Mariona Taulé

Professora del Departament de Lingüística General (Universitat de Barcelona)
Consultora dels Estudis d'Humanitats i Filologia (UOC)
mtaule@pcb.ub.es



M. Antònia Martí

Professora del Departament de Lingüística General (Universitat de Barcelona)
amarti@lingua.fil.ub.es

Resum:

El propòsit d'aquest article és presentar de manera general el certamen internacional SENSEVAL, l'objectiu bàsic del qual és l'avaluació objectiva de tècniques, mètodes i sistemes de desambiguació semàntica automàtica. Amb aquesta finalitat, primer es fa una presentació global de SENSEVAL, els seus objectius i l'organització; a continuació s'exposa la metodologia d'avaluació desenvolupada, que consisteix en la definició de tasques i la preparació dels recursos lingüístics necessaris, i, finalment, s'analitzen les repercussions en els àmbits lingüístic i lexicogràfic que hi ha al darrere d'aquesta metodologia. Per a exemplificar la metodologia ens basarem en els recursos desenvolupats per al SENSEVAL-2 de l'espanyol.

1. Introducció

Un dels problemes centrals i més difícils de tractar en les diferents aplicacions de processament del llenguatge natural (d'ara endavant, PLN) és, sens dubte, l'ambigüitat, una propietat intrínseca i característica del llenguatge. Des d'un punt de vista lingüístic, s'entén per ambigüitat el fenomen que es dona quan una mateixa forma lingüística es pot interpretar de més d'una manera diferent. Quan l'ambigüitat és motivada pels múltiples significats d'una mateixa paraula aleshores parlem d'ambigüitat semàntica i, en concret, d'ambigüitat lèxica. Aquest és el cas, per exemple, de la paraula *partit*, que pot significar bàsicament tres coses diferents: a) un nom que designa una organització política ("Militava en un **partit** polític"), b) un nom que designa una prova esportiva ("El **partit** de futbol va acabar amb un empat"), i, finalment, c) el participi del verb *partir*, que indica que alguna cosa està dividida o separada ("El vaixell ha quedat **partit** en dues parts"). Com es pot observar, segons la categoria i, especialment, segons el context en què apareix, la paraula adquireix un significat o un altre. En aquest sentit, no es pot tractar el significat de les paraules d'una manera aïllada, perquè habitualment és el context el que perfila, concreta o modula el seu significat, i és a partir d'aquest que es pot obtenir el sentit ple o més apropiat. Quan parlem de context ens referim tant al context sintàctic immediat, a l'oració en la qual ocorre la paraula, a les paraules amb què es combina, com al context més ampli, per exemple el paràgraf o text en el qual es troba la

paraula.

L'ambigüitat lèxica no representa, en principi, i llevat de casos molt concrets, cap problema per als humans, però sí, en canvi, resulta d'una extraordinària complexitat per als sistemes de tractament automàtic del llenguatge i, especialment, quan el que s'ha de determinar és el significat d'una expressió lingüística, és a dir, d'una determinada paraula o seqüència de paraules. La desambiguació semàntica no és una tasca gens senzilla. Per a un programa, qualsevol forma (o paraula) a la qual es pugui assignar més d'una interpretació, sigui morfosintàctica o semàntica, és una entitat en principi ambigua, perquè el programa haurà de decidir quina de les possibles interpretacions és la més adequada, això sí, sempre tenint en compte el context en què apareix. Si no partim del context, el programa o mòdul de desambiguació no té cap sentit. Doncs bé, els desambiguadors morfosintàctics (*taggers*) i els sistemes de desambiguació semàntica (*Word Sense Disambiguation*) són els que s'encarreguen de decidir i assignar la correcta interpretació morfosintàctica o semàntica, respectivament, d'una paraula concreta. Quan la desambiguació es porta a terme en l'àmbit de la categoria lèxica s'utilitza un *tagger*, i quan la desambiguació implica determinar el sentit o significat més probable d'una determinada paraula llavors s'utilitza un desambiguador semàntic (o lèxic). En general, el procés de desambiguació semàntica es realitza una vegada analitzat i desambigat morfosintàcticament el text, ja que d'aquesta manera es redueix l'ambigüitat semàntica al pla de categoria. Per exemple, en el cas de la paraula *partit*, el desambiguador morfosintàctic distingeix entre les formes nominals i les formes de participi i, en conseqüència, redueix l'ambigüitat d'aquesta paraula a les ocurrències amb categoria nom. Seguint amb l'exemple, el desambiguador semàntic només haurà de decidir, en el cas de les formes nominals, entre dos sentits possibles ("partit polític" o "prova esportiva") i no tres. Per tant, l'aplicació prèvia del desambiguador morfosintàctic redueix considerablement les possibilitats d'assignació semàntica de sentits.

Un altre aspecte que cal tenir en compte en el procés de desambiguació semàntica és el fet que per a interpretar el significat d'una oració és necessari determinar, prèviament, el significat de cadascuna de les paraules que la formen. Per tant, per a resoldre la semàntica de l'oració s'ha de resoldre primer l'ambigüitat lèxica.

El problema de la resolució de l'ambigüitat semàntica (o lèxica) és un problema subjacent en moltes de les aplicacions de PLN, especialment en els sistemes de traducció automàtica i de recuperació de la informació, un problema que no és nou, que es va plantejar des de l'origen del tractament automàtic del llenguatge (Ide i Véronis, 1998). En general, tots aquests sistemes incorporen un mòdul, més o menys efectiu, de desambiguació semàntica per a poder obtenir resultats d'anàlisis òptims. Ara bé, recentment l'interès per aquest tipus de tècniques o de sistemes s'ha incrementat considerablement, cosa que és possible que es degui en gran part al fet que hi hagi cada vegada una major disponibilitat de corpus de grans dimensions, que permeten aplicar mètodes estadístics per a tractar la desambiguació semàntica i que, normalment, resulten molt efectius. Els corpus, col·leccions de text en format electrònic, es converteixen en extraordinaris dipòsits d'informació a partir de la qual es poden trobar, obtenir i, per tant, aprendre els múltiples contextos en els quals pot aparèixer una determinada paraula, de manera que esdevenen una font d'informació fonamental per als sistemes de desambiguació.

La tendència general és el desenvolupament de sistemes de desambiguació més genèrics, és a dir, independents de l'aplicació, amb l'objectiu centrat a resoldre l'assignació de sentits. Aquesta proliferació de sistemes ha originat la creació del concurs SENSEVAL (acrònim de *Sense Evaluation*), que permet comparar i avaluar els resultats obtinguts pels diversos sistemes de desambiguació semàntica automàtica d'una manera objectiva. Per a poder fer aquesta avaluació objectivament cal desenvolupar una metodologia d'avaluació que permeti comparar en igualtat de condicions els diferents sistemes. Una metodologia que implica, d'alguna manera, una revisió exhaustiva de la definició de sentit i de com s'han de discriminar o distingir els sentits bàsics de les paraules. És a dir, una metodologia que també ha tingut una repercussió important en plantejaments fonamentals de lexicografia computacional i de lingüística en general.

L'objectiu d'aquest article és presentar la metodologia desenvolupada a SENSEVAL, explicar per què s'ha convertit en el referent o el marc imprescindible per als sistemes de desambiguació semàntica i reflexionar sobre les repercussions en els àmbits lingüístic i lexicogràfic que hi ha al darrere d'aquesta metodologia. Per a presentar-ho hem organitzat el text en tres parts: a la primera, es fa una presentació global de SENSEVAL, els seus objectius i l'organització; a continuació, s'exposa la metodologia d'avaluació desenvolupada, i, finalment, les repercussions que ha tingut en la lingüística i la lexicografia. Per a exemplificar la metodologia ens basarem en els recursos desenvolupats per al SENSEVAL-2 de l'espanyol.^[1]

2. Què és SENSEVAL?

SENSEVAL^[2] és un certamen internacional sobre desambiguació semàntica automàtica (Kilgarrif i Palmer, 2000; Edmonds i Cotton, 2001), que se celebra bianualment des de l'any 1998 i l'objectiu principal del qual és l'avaluació objectiva de sistemes, mètodes i tècniques que determinen automàticament el significat d'una paraula en un context. L'avaluació es fa per a diferents paraules, diferents varietats lingüístiques i diferents llengües. En aquest marc d'intercanvi científic no tan sols es tracta d'explorar, de comprovar i comparar els aspectes tant teòrics com tècnics dels diversos programes dissenyats per a la desambiguació semàntica de sentits (*Word Sense Disambiguation*, WSD), sinó que també s'ha generat, gràcies a aquesta iniciativa, tota una reflexió sobre el lèxic, sobre com s'ha de tractar, analitzar i, en definitiva, representar el significat de les paraules, i, per tant, es planteja tota una problemàtica entorn de la polisèmia des d'un punt de vista experimental molt interessant i nou.

El primer certamen SENSEVAL^[3] es va celebrar el 1998 i les llengües participants van ser l'anglès, el francès i l'italià. La metodologia desenvolupada permetia avaluar els sistemes de desambiguació exclusivament partint d'una mostra lèxica, és a dir, s'havia de determinar automàticament el sentit d'una sola paraula en un context determinat. Amb SENSEVAL-2,^[4] les llengües participants es van incrementar fins a un total de dotze (anglès, francès, italià, espanyol, basc, danès, suec, holandès, estonià, txec, xinès i japonès), de la mateixa manera que també van augmentar els tipus de tasques possibles en què podien concursar els diferents participants. En aquest sentit, es van dissenyar tres tipus de tasques que bàsicament es diferenciaven entre elles pel nombre de paraules que els programes havien de desambiguar. Amb la finalitat d'entendre millor la metodologia que se segueix a SENSEVAL, presentem a continuació de manera més detallada en què consisteix cadascuna d'aquestes tasques i els recursos desenvolupats per a poder fer-les.

3. Metodologia

Els diferents algoritmes i sistemes de desambiguació semàntica que participen a SENSEVAL tenen l'objectiu de resoldre el problema de determinar automàticament el sentit de les paraules sotmeses a examen, en un context determinat, és a dir, decidir i assignar a cada paraula el significat més probable segons el context sintàctic en què apareix.

Per a poder fer una avaluació objectiva, primerament és necessari desenvolupar una metodologia d'avaluació en què es defineixin explícitament les diverses tasques en les quals poden participar els diferents programes i a partir de les quals s'han de comparar i avaluar els resultats. En segon lloc, s'han de proporcionar als participants les dades necessàries, corpus i lèxics, per a poder realitzar de manera objectiva l'avaluació. S'entén que es tracta de les mateixes dades per a tots els sistemes que participen en la desambiguació d'una mateixa llengua. Finalment, també s'ha d'elaborar un programa que comptabilitzi, compari i avaluï els

1. Els recursos desenvolupats per a la realització de SENSEVAL-2 han estat possibles gràcies a la col·laboració desinteressada dels grups de recerca del Centre de Llenguatge i Computació de la Universitat de Barcelona (CLIC-UB), el centre de Tècniques i Aplicacions del Llenguatge i la Parla (TALP) de la Universitat Politècnica de Catalunya (UPC) i el Grup de Processament de Llenguatge Natural (GPLN) de la Universitat Nacional d'Educació a Distància (UNED). Aquests tres grups s'han encarregat de l'organització del SENSEVAL-2 espanyol.

2. L'organització de SENSEVAL^[url1] es porta a terme amb la col·laboració desinteressada de diferents grups de recerca.

3. El primer exercici de SENSEVAL, i de fet també el seu homònim ROMANSEVAL (Véronis, 1998; Calzolari i Corazzari, 2000), va tenir lloc el 1998, sota la supervisió del grup SIGLEX, i va culminar en un taller (*workshop*) a Hertsmonceux Castle, Anglaterra. SIGLEX^[url2] (Special Interest Group of the Association for Computational Linguistics) proporciona un marc de referència per a la recerca en lexicografia, semàntica lèxica computacional i altres àrees afins.

4. SENSEVAL-2 es va organitzar el 2000-2001 amb el suport d'EURALEX, ELSNET, EPSRC i ELRA, i va tenir lloc a Tolosa juntament amb el congrés de l'ACL-2001.

resultats que han obtingut els diferents participants (Melamed i Resnik, 2000).

3.1. Tipus de tasques

A SENSEVAL-2^[5] es van dissenyar tres tipus de tasques per a avaluar els sistemes de desambiguació: tasca basada en una mostra lèxica (*lexical sample task*), tasca lèxica completa (*all-words task*) i tasca de traducció (*translation task*).

La tasca basada en la mostra lèxica només es fixa en una sola paraula per frase, mentre que a la tasca lèxica completa els sistemes han de desambiguar semànticament totes les paraules, tret de les funcionals (conjuncions, preposicions, articles, etc.), que apareixen a les frases dels corpus seleccionats. La tasca de traducció, en la qual només ha participat la llengua japonesa, és de fet un subtipus de mostra lèxica perquè només s'ha de desambiguar una sola paraula; la diferència és que el sentit de la paraula es defineix d'acord amb la seva traducció.

En definitiva, en la tasca lèxica completa s'avalua la capacitat del sistema automàtic de desambiguar cadascuna de les paraules (o ocurrencies) que apareixen al corpus que es proporciona. En les tasques basades en una mostra lèxica s'avaluen els sistemes segons la desambiguació de les diferents ocurrencies d'una mostra de paraules prèviament seleccionades en el context d'una frase.

A la taula següent es mostren de manera esquemàtica les tasques en què han participat les diferents llengües i el nombre d'equips participants (entre parèntesis):

TASQUES	
Tasca basada en una mostra lèxica	Espanyol (5), basc (2), anglès (15), italià (2), japonès (3), coreà (2), suec (5)
Tasca de traducció	Japonès (8)
Tasca lèxica completa	Txec (1), estonià (2), holandès ⁱ , anglès (12)

ⁱ En el cas de la llengua holandesa es van preparar les dades, però no van estar disponibles per al concurs

Taula 1. Tasques en què participen les diverses llengües

Per a poder dur a terme aquestes tasques cal disposar de dos tipus bàsics de dades: un diccionari (o lèxic) i un corpus etiquetat manualment que servirà de *gold-standard* i a partir del qual es crearà el corpus d'entrenament i el corpus d'avaluació.^[6] Tots els sistemes que hi concuren han d'analitzar semànticament el mateix corpus i, en conseqüència, és necessari que cada llengua participant disposi d'aquest corpus i d'un diccionari en el qual s'inclouin els diferents significats de les paraules que s'han de desambiguar.

5. A partir d'ara s'exposa la metodologia desenvolupada i les dades i resultats obtinguts a SENSEVAL-2 perquè comprèn més llengües i més tasques. De fet, la metodologia del primer concurs SENSEVAL queda inclosa al segon.

6. A Rigau i altres (2001) i a Taulé (2002) es dona una descripció completa dels recursos lingüístics elaborats per a SENSEVAL-2.

3.1.1. Lèxic

En el cas de la tasca basada en una mostra lèxica, el lèxic que es proporciona té el requisit d'haver de contenir el conjunt de paraules seleccionades per a l'exercici.^[7] Per a la tasca lèxica completa evidentment es necessita un lèxic extens. En tot cas es tracta de lèxics disponibles en format electrònic.

A continuació, es donen com a exemple dues entrades, *grano* i *popular*, del lèxic^[8] elaborat expressament perquè la llengua espanyola pogués participar en la tasca de mostra lèxica.

grano#NCMS#1#Pequeño bulto o forúnculo que sale en la piel: la nariz roja y con muchos granos

#SIN:forúnculo

#COL:grano de pus

#SYNSET:08681778n/08677034n#

grano#NCMS#2#Pequeña partícula o elemento que forma parte de un todo: doce granos de uva; granos de polvo

#SIN:?

#COL:grano de arena, grano de uva, grano de azúcar

#SYNSET:05734433n#

grano#NCMS#3#Semilla o fruto de los cereales o de otras plantas: granos de cereal; grano de maíz

#SIN:?

#COL:grano tostado, grano de trigo, grano de arroz, grano de café, grano de maíz

#SYNSET:07123531n/07123713n#

LOCUCIONES:

ir al grano#

popular#AQCS#1#Relativo al pueblo: música popular; tradiciones populares

#SIN:folclórico

#ANT:?

#COL:?

#SYNSET:00537687a/00363173a/05171170a/02139261a/00834378a#

popular#AQCS#2#Que gusta, que es muy conocido: un cantante popular; un actor muy popular

#SIN:famoso

#ANT:desconocido, anónimo

#COL:?

#SYNSET:01386309a/01047611a#

popular#AQCS#3#Relativo al partido popular: dirigente del partido popular; un diputado popular

7. Aquest és el cas del SENSEVAL espanyol, que no disposava d'un lèxic complet, per la qual cosa se'n va fer un de parcial dissenyat especialment per al concurs.

8. El lèxic (o minidiccionari) del SENSEVAL espanyol ha estat elaborat íntegrament pel grup de lingüistes del Centre de Llenguatge i Computació de la Universitat de Barcelona (CLiC-UB).

#SIN:?

#ANT:?

#COL:dirigente popular, sede popular, partido popular

#SYSNET:?#

LOCUCIONES:

tribunal popular#

La informació que contenen les entrades d'aquest lèxic^[9] (d'ara endavant "Minidir"), especialment confeccionat per a SENSEVAL i, per tant, amb una clara orientació per a ser utilitzat per un sistema de desambiguació semàntica, s'estructura en set (en el cas de les entrades nominals i verbals) o vuit (en les entrades adjectivals) camps diferents: el lema que identifica l'entrada, la categoria lèxica,^[10] el nombre d'accepció, la definició seguida opcionalment d'un exemple, els sinònims, els antònims en el cas de les entrades adjectives, una llista de les col·locacions més freqüents i el *synset* o *synsets*^[11] corresponents d'EuroWordNet.^[12] Aquesta informació la té associada cadascuna de les accepcions de l'entrada i al final també es proposa una llista de locucions.

Un dels aspectes més problemàtics que s'han plantejat en l'elaboració del "Minidir" és l'assignació del *synset* corresponent a cada un dels sentits de les paraules seleccionades, és a dir, el fet de compatibilitzar la informació del diccionari amb la de la base de coneixement EuroWordNet. Al "Minidir" s'hi han mirat d'incloure només els significats bàsics, amb la qual cosa s'ha obtingut un lèxic que presenta poca granularitat de sentits, una mitjana de quatre sentits per a les entrades nominals i adjectives i de sis per als verbs. Tanmateix, a EuroWordNet les paraules representades presenten una gran granularitat, és a dir, especificitat, de sentits. Per això, trobem que al "Minidir" una mateixa accepció es pot correspondre amb més d'un *synset* de la base de coneixement (per exemple: "grano_1: Pequeño bulto o forúnculo que sale en la piel" es correspon amb dos *synsets* diferents -08681778 n/08677034n-), i així es fa constar en els sentits corresponents.

A més dels lexicons proporcionats també s'ha fet servir la versió 1.7 de les bases de coneixement lèxic-semàntic WordNet,^[13] per a les tasques de l'anglès, i EuroWordNet, per a les tasques de l'espanyol, l'italià i l'estonià. Per aquest motiu, s'han associat les definicions de cada paraula del lèxic amb el *synset* corresponent per a tenir ambdues fonts relacionades.

A continuació es fa una llista de les paraules seleccionades per a realitzar la tasca basada en una mostra lèxica de l'espanyol:

NOMS
<i>arte, autoridad, bomba, canal, circuito, corazón, corona, gracia, grano, hermano, masa, naturaleza, operación, órgano, partido, pasaje, programa, tabla</i>
VERBS
<i>actuar, apoyar, apuntar, clavar, conducir, copiar, coronar, explotar, saltar, tocar, tratar, usar, vencer</i>
ADJECTIUS
<i>brillante, claro, ciego, local, natural, popular, simple, verde, vital</i>

Taula 2. Paraules seleccionades per al SENSEVAL espanyol

9. El diccionari només inclou les quaranta paraules seleccionades per a la prova (18 noms, 9 adjectius i 13 verbs).

10. La categoria lèxica (nom, adjectiu o verb) es representa amb etiquetes *parole* (Eagles, programa Eureka 1989-1995) abreujades.

11. En el model de WordNet els conceptes es defineixen per conjunts de sinonímia, anomenats *synsets*, connectats entre ells per relacions semàntiques (hiponímia, hiperonímia, antonímia, meronímia, etc.).

12. A Vossen (1999), hi ha una descripció completa d'EuroWordNet^[tur3]. Lloc web d'EuroWordNet en espanyol, català i basc: <<http://nipadio.lsi.upc.es/wei.html>[tur4]>.

13. A Miller i Fellbaum (1991) hi ha una descripció completa de WordNet^[tur5].

3.1.2. Corpus

El segon recurs lingüístic necessari per a dur a terme l'avaluació dels sistemes és la creació del corpus, l'anomenat "*gold-standard*", a partir del qual s'obté el corpus d'entrenament i el corpus d'avaluació. Per a poder avaluar i comparar els resultats dels diferents sistemes cal disposar d'aquest corpus desambiguat manualment, de manera que sigui possible de fer-lo servir com a referència a partir de la qual es puguin comparar els resultats obtinguts pels sistemes participants.

En el cas de les tasques basades en una mostra lèxica, l'elaboració del *gold-standard* es fa a partir de les paraules preseleccionades per a l'avaluació, amb la qual cosa s'obté com a context mínim l'oració en la qual apareixen aquestes paraules. Seguint les indicacions proposades per l'organització de SENSEVAL (Edmonds i Cotton, 2001), es mira de proporcionar com a mínim quinze ocurrences diferents de cada sentit o accepció d'aquestes paraules.^[14]

Una vegada seleccionat el corpus es fa l'anotació semàntica, és a dir, es porta a terme la seva desambiguació, que consisteix a assignar a cada paraula el sentit més apropiat del diccionari. Aquest procés es realitza manualment, i ho fa més d'una persona, procurant que s'esdevingui un 90% d'acord final entre els diferents anotadors. Una part del *gold-standard* es lliura als participants un mes abans del concurs perquè els seus sistemes de desambiguació el facin servir com a corpus d'entrenament, mentre que la resta s'utilitza com a corpus d'avaluació per al concurs, esborrant prèviament les anotacions semàntiques. Els sistemes disposen d'una setmana per a lliurar els resultats obtinguts pels seus sistemes a l'organització, que els compararà amb les anotacions proposades al *gold-standard* pels anotadors humans.

S'ha d'assenyalar que també es distingeix entre els sistemes supervisats, és a dir, els que utilitzen el corpus d'entrenament prèviament anotat, i els sistemes no supervisats, que no el fan servir perquè empen un altre tipus de fonts de coneixement, com ara diccionaris o altres corpus disponibles. En el cas de l'espanyol, tots els sistemes que es van presentar eren supervisats i requerien, per tant, el corpus d'entrenament.

A continuació es mostra l'entrada verbal *explotar* i una part del *gold-standard* espanyol en què apareix aquest verb desambiguat semànticament.

- Entrada lèxica d'*explotar*:

explotar#VM#1#Obtener beneficio de algo mediante el trabajo: explotar una mina

#SIN:?

#COL:?

#SYNSET:00660933v/00661466v#

explotar#VM#2#Hacer trabajar a alguien de forma abusiva en beneficio propio: explotaba a sus empleados

#SIN:abusar, aprovecharse

#COL:?

#SYNSET:01363660v#

explotar#VM#3#Valerse de una situación o de una relación en beneficio propio: explotaba sus encantos para obtener trabajo

#SIN:aprovecharse

14. Hi ha casos en què no és possible proporcionar quinze ocurrences per sentit pel fet de no trobar més exemples als corpus a partir dels quals s'obté el *gold-standard*.

#COL:?

#SYNSET:00659593v/01303135v#

explotar#VM#4#Hacer explosión una cosa, en especial un artefacto: una bomba muy potente explotó cerca del mercado

#SIN:estallar

#COL:?

#SYNSET:00185045v/00185462v/00643046v#

explotar#VM#5#Manifestar o surgir de repente un sentimiento o una situación latente: el presidente explotó indignado por las declaraciones

#SIN:estallar

#COL:?

#SYNSET:00185244v/00404923v#

- Mostra del *gold-standard* de l'espanyol:

La empresa Boliden, que **explota** la mina de Aznalcóllar, dice ahora_que aplaza la reapertura de la instalación. #1

El suceso más grave tuvo lugar en Bilbao, donde un artefacto casero **explotó** ante la puerta del domicilio de Dimas_Sañudo, candidato del PSE a la alcaldía. #4

Se **explota** el filón de su etapa en Ercros y de lo que él mismo ha definido como optimización fiscal: el uso de una sociedad patrimonial para pagar menos al fisco. #3

Tamames **explotó**: "¡Poner a Clinton de modelo de virtud! #5

Un pequeño artefacto **explotó** ayer en el buzón del domicilio de la concejal del PP en Zarauz_Lucía_Peralta, sin que se registraran heridos. #4

Según esas fuentes, la mujer, identificada por la policía como Esmá_Yurdakul, hizo **explotar** los artefactos que llevaba encima cuando pasó al_lado_de dos vehículos policiales estacionados en el centro de la plaza Taksim, una de las más frecuentadas de la capital. #4

Dave_Morris y Helen_Steel fueron condenados en 1997 a pagar 15 millones de pesetas por repartir panfletos en los que afirmaban que la empresa produce alimentos perniciosos para la salud, que maltrata a los animales y que **explota** a sus empleados. #2

El primer ministro italiano, Massimo_D'Alema , aseguró ayer que "Albania corre el riesgo de **explotar**" , como consecuencia de la avalancha de refugiados procedentes de Kosovo. #5

La firma que **explota** la ruta a Barcelona acumula quejas de los usuarios y sanciones. #1

Los inmigrantes eran **explotados** por el grupo como peones en unas condiciones de absoluta esclavitud. #2

Constructoras más que dudosas se han valido de obreros sin papeles, a los que se ha **explotado** de manera vergonzosa. #2

El fuego hizo que **explotara** una de las lunas de cristal del cajero y los vidrios hirieron al trabajador Carlos_Elizalde, produciéndole heridas en la cara, los brazos y las piernas. #4

El corpus per al SENSEVAL espanyol s'ha elaborat a partir de dues fonts diferents: a) *El Periódico*,^[15] un corpus que recull les notícies de l'any 2000 extretes d'aquest diari; es tracta, per tant, d'una recopilació de textos en llenguatge estàndard i de temàtica general. b) *Lexesp*,^[16] un corpus equilibrat de 5,5 milions de paraules, que inclou textos de diversos temes (científics, econòmics, jurídics, literaris, periodístics, etc.) i estils (assaig, novel·la, etc.) que recullen diversos registres de la llengua (estàndard, tècnic, culte, etc.). Com que són dos corpus analitzats i desambiguats morfològicament,^[17] és possible fer de manera ràpida la cerca automàtica de les frases que contenen les paraules seleccionades per a l'exercici (una cerca combinada de lema i categoria a fi d'extreure'n totes les formes en les quals apareix aquesta paraula als corpus). L'elaboració del *gold-standard* es va realitzar en dues fases: primer, es va fer l'etiquetatge semàntic del corpus i, en una segona fase, es van comparar els resultats aportats pels diferents equips d'annotadors;^[18] en aquells casos en què no es produïa acord en l'assignació de sentits (en un 46% de les ocurrències) un sol anotador intervenia com a àrbitre per a assignar l'etiqueta definitiva.

Quant als sistemes participants, en la tasca lèxica de l'espanyol es van presentar un total de dotze sistemes de cinc equips diferents: Universitat d'Alacant (UA), Universitat John Hopkins (JHU), Universitat de Stanford (SU), Universitat de Maryland (UMD) i Universitat de Manitoba (que va presentar vuit algorismes diferents de desambiguació, D6-D10, dX, dY i dZ). El sistema que va obtenir millors resultats, és a dir, que s'aproximava més als resultats obtinguts manualment, va ser el sistema d'aprenentatge supervisat de la Universitat John Hopkins. Aquest sistema va obtenir millors resultats en 12 de les 39 paraules seleccionades, i també va ser el millor en la desambiguació de noms i verbs, però no en el cas dels adjectius, per als quals la Universitat de Stanford va obtenir resultats més bons. En general, en tots els sistemes, la desambiguació de noms i adjectius s'acostava més als resultats etiquetats del *gold-standard* que la dels verbs. A la taula 3 es mostren els percentatges d'acord obtinguts pels diferents sistemes participants.

Sistemes	UA	SU	JHU	UMD	d6	d7	d8	d9	d10	dX	dY	dZ
Acord	0,51	0,63	0,65	0,61	0,55	0,57	0,59	0,53	0,59	0,55	0,51	0,57
Kappa	0,20	0,34	0,47	0,20	0,13	0,19	0,23	0,06	0,24	0,15	-0,03	0,15

Taula 3. Percentatges d'acord i mesura *Kappa*

Com es pot observar el sistema presentat per la JHU és el que té una mitjana d'encert superior: 0,65 o 0,47 si hi apliquem la mesura *Kappa*.

Per a veure els resultats obtinguts pels diferents participants més detalladament, com també els recursos desenvolupats per a l'exercici, consulteu la pàgina oficial de SENSEVAL^[ur16].

4. Repercussions de SENSEVAL en els estudis de semàntica lèxica

La metodologia desenvolupada a SENSEVAL orientada principalment a l'avaluació i la comparació de sistemes automàtics de desambiguació semàntica s'ha centrat fonamentalment en la valoració dels algorismes d'aquests sistemes. Aquests algorismes s'han avaluat independentment de la qualitat lingüística i lexicogràfica dels recursos lingüístics, tant de les fonts lèxiques com dels corpus anotats a mà (o *gold-standard*) utilitzats. En aquest context, s'ha generat una discussió paral·lela entorn de qüestions bàsiques de lexicografia i semàntica, que

15. El corpus d'*El Periódico* ha estat cedit pel projecte HERMES (CICYT TIC2000-0335-C03-02).

16. El corpus *Lexesp* ha estat proporcionat pel projecte Lexesp-III (DGICYT APC 99-0105).

17. Per a l'anàlisi morfològica s'ha fet servir l'analitzador morfològic MACO+ (Carmona i altres, 1998) i per a la desambiguació morfològica el *tagger* RELAX (Padró, 1998).

18. Els equips de desambiguació eren constituïts per l'equip de lingüistes del CLiC (UB) i els equips, integrats bàsicament per informàtics, del TALP (UPC) i del GPLN (UNED).

són a la base de tot el procés de desambiguació. Els punts en els quals se centra actualment el debat són els següents:

- La validesa del procés d'etiquetatge, tant de l'etiquetatge manual per als corpus d'aprenentatge com de l'etiquetatge automàtic, en els processos de desambiguació semàntica automàtica.
- L'objectivitat en els processos d'avaluació dels sistemes de desambiguació semàntica automàtica.
- La qualitat de les fonts d'informació i la seva adequació per a les tasques de desambiguació semàntica de sentits.
- La mateixa existència dels sentits, la seva descripció i representació.

Una de les primeres qüestions que es poden plantejar és si és possible o no la desambiguació semàntica, tant manual com automàtica. Si els humans actuem de manera poc sistemàtica en aquesta tasca, llavors el valor dels corpus etiquetats semànticament és qüestionable, i també ho serà el desenvolupament de sistemes que facin la desambiguació semàntica de manera automàtica. Si bé és cert que les màquines realitzen unes determinades tasques millor que els humans, això no és així quan hi ha implicat el significat, el context i, en general, la intel·ligència. En aquests casos, es pot considerar que la manera en què ho fa un humà és el nivell màxim de qualitat al qual es pot aspirar. Així, la qualitat dels sistemes de desambiguació semàntica automàtica s'avalua comparant els seus resultats amb corpus etiquetats manualment (*gold-standard*), cosa que implica analitzar com duem a terme aquesta tasca els humans i la validesa dels resultats obtinguts.

Arran de les recerques fetes en l'àrea de la desambiguació semàntica automàtica i dels experiments orientats a avaluar la qualitat de les fonts i la validesa del procés d' anotació (Véronis, 2002; Krishnamurthy i Nicholls, 1998), s'ha posat de manifest l'escàs acord entre els anotadors humans i la poca adequació de les fonts de coneixement lèxic per a la realització de la desambiguació semàntica automàtica, cosa que ha portat no tan sols al replantejament del mateix concepte de sentit, de com s'ha de descriure i representar, sinó que també es qüestiona la consistència dels mètodes de discriminació de sentits. Alguns experiments anteriors s'havien centrat a analitzar com els humans fem aquesta tasca (Ahlsvede, 1995; Ahlsvede i Lorand, 1993; Fellbaum, Grabowsky i Landes, 1998), encara que aquests estudis o bé són poc sistemàtics o tracten un nombre molt reduït de paraules. Véronis (2002) proposa un experiment amb l'objectiu de mostrar la falta d'acord en l'anotació manual de corpus que conclou amb un diagnòstic sobre les raons del problema: la no-adequació dels diccionaris d'ús comú ni de les antologies existents per a la realització de la desambiguació semàntica automàtica.

Una altra de les conseqüències d'aquest tipus de recerques és el replantejament d'algunes qüestions que afecten els fonaments de la polisèmia. S'ha evidenciat la necessitat de definir unes bases teòriques i metodològiques per al tractament dels sentits, tant referent a la seva discriminació o identificació com pel que fa als criteris de granularitat (especificitat) amb què s'han de tractar per a tasques de desambiguació semàntica automàtica. En conseqüència, s'han qüestionat alguns dels procediments del treball lexicogràfic, especialment el recurs de la intuïció per a la distinció de sentits, per a obrir-lo a noves orientacions i tipus de coneixement (Kilgarrif, 1998; Miller i Leacock, 2000) com la informació sintàctica, col·locativa i exemples extrets de corpus. En aquesta línia val la pena destacar els treballs de Hanks (2000), Palmer (1998), Mana i Corazzari (2000) i Véronis (2002).

Una altra de les línies de treball recent es fonamenta en la idea que la comparació entre les llengües és útil per a la desambiguació semàntica automàtica. Es parteix de la hipòtesi que la correspondència entre les paraules i els sentits varia de manera significativa d'una llengua a una altra (Ide, 2000). Resnik i Yarowsky (1997) suggereixen que, per als propòsits de la desambiguació semàntica automàtica, els múltiples sentits d'una paraula es podrien determinar si es consideressin només les distincions de sentit lexicalitzades entre llengües

cross-linguistically).

Són moltes les crítiques que han sorgit en contra de l'exclusiva representació dels sentits en forma de definicions d'un diccionari, tant per la manca d'informació sobre les preferències lèxiques i les estructures sintàctiques que accepten, com per la imprecisió del llenguatge utilitzat i el mètode d'anàlisi mateix, la introspecció. Es tracta de veure fins a quin punt les tècniques i mètodes que es proposen des del processament del llenguatge natural es poden aplicar per a millorar les fonts de coneixement existents, proposar noves fonts d'informació i aportar criteris per a la discriminació i desambiguació de sentits.

Respecte a això, SENSEVAL s'ha convertit en un nou fòrum de discussió, d'anàlisi i d'intercanvi científic en el qual tenen cabuda tots els aspectes vinculats amb la resolució de l'ambigüitat semàntica, sigui des d'un punt de vista més tècnic, com és el cas de les estratègies i mètodes proposats per al seu tractament automàtic, sigui des d'un punt de vista més lingüístic, que posa en evidència la dificultat que implica la delimitació de sentits i la necessitat d'un estudi amb profunditat sobre la seva discriminació, és a dir, sobre la polisèmia i sobre els criteris lingüístics que en permetin la delimitació d'una manera objectiva, rigorosa i adequada per als sistemes de desambiguació semàntica automàtica.

Llista d'URL:

- [url1]:<http://www.senseval.org/>
- [url2]:<http://www.siglex.org/>
- [url3]:<http://www.illc.uva.nl/EuroWordNet/>
- [url4]:<http://nipadio.lsi.upc.es/wei.html>
- [url5]:<http://www.cogsci.princeton.edu/~wn/>
- [url6]:<http://www.senseval.org>

Bibliografia:

- AHLWEDE, T.E. (1995). "Word sense disambiguation by humans informants". A: *Proceedings of the 6th Midwest Artificial Intelligence and Cognitive Society Conference* (abril 1995: Carbondale, Illinois), pàg. 73-78.
- AHLWEDE, T.E.; LORAND, D. (1993). "The ambiguity questionnaire: a study of lexical disambiguation by human informants". A: *Proceedings of the 6th Midwest Artificial Intelligence and Cognitive Society Conference* (Chesterton, Indiana), pàg. 21-25.
- CARMONA, J.; CERVELL, S.; MÀRQUEZ, L.; MARTÍ, M.A.; PADRÓ, L.; PLACER, R.; RODRÍGUEZ, H.; TAULÉ, M.; TURMO, J. (1998). "An environment for morphosyntactic processing of unrestricted spanish text". A: *Proceedings of the First International Conference on Language Resources and Evaluation, LREC* (Granada, Espanya).
- EDMONDS, P.H.; COTTON, S. (2001). "SENSEVAL-2 overview". A: *Proceedings of SENSEVAL-2, Second International Workshop on Evaluating Word Sense Disambiguation Systems, 39th Annual Meeting of the ACL* (França).
- FELLBAUM, C.; GRABOWSKY, L. (1998). "Performance and confidence in a semantic annotation task". A: FELLBAUM, C. (ed.). *WordNet: An electronic database*. Cambridge (Massachusetts): MIT Press, pàg. 217-237.
- GALE, W.A.; CHURCH, K.W.; YAROWSKY, D. (1993). "A method for disambiguating word senses in a large corpus". *Computer and the Humanities*. Núm. 26, pàg. 415-439.
- HANKS, P. (2000). "Do word meanings exist?". *Computers and the Humanities* (número especial: *Evaluating Word Sense Disambiguation Programs*). Vol. 1-2, núm. 34, pàg. 205-215.
- IDE, N. (2000). "Cross-lingual sense determination: Can it work?". *Computers and the Humanities* (número especial: *Evaluating Word Sense Disambiguation Programs*). Vol. 1-2,

núm. 34, pàg. 223-234.

IDE, N.; VÉRONIS, J. (1998). "Introduction to the special issue on word sense disambiguation: the state of the art". *Computational Linguistics*. Vol. 1, núm. 24, pàg. 1-40.

KILGARRIFF, A. (1998). "SENSEVAL: An exercise in Evaluating Word Sense Disambiguation Programs". LREC-1998.

KILGARRIF, A.; PALMER, M. (2000). (editors convidats). "Special issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs". *Computers and the Humanities*. Vol. 1-2, núm. 34.

KILGARRIF, A.; ROSENZWIG, J. (2000). "Framework and results for english SENSEVAL". *Computers and the Humanities*. Vol. 1-2, núm. 34, pàg. 15-48.

KRISHNAMURTHY, R.; NICHOLLS, D. (1998). "Peeling an onion: the lexicographer's experience of manual sense tagging". A: *SENSEVAL Workshop (2-4 de setembre de 1998: Herstmonceux Castle, Sussex, Anglaterra)*.

MANA, N.; CORAZZARI, O. (2001). "The lexico-semantic annotation of an italian treebank" [en línia]. <<http://www.tcc.itc.it/publications/2001/>>.

MELAMED, D.; RESNIK, P.H. (2000). "Tagger evaluation given hierarchical tag sets". *Computers and the Humanities*. Vol. 1-2, núm. 34.

MILLER, G.A.; FELLBAUM, C. (1991, desembre). "Semantic networks of English". *Cognition*. Núm. 41, pàg. 197-229.

MILLER, G.A.; LEACOCK, C. (2000). "Lexical representations for sentence processing". A: *Polysemy, theoretical and computational approaches*. Oxford University Press.

PADRÓ, L. (1998). *An hybrid environment for syntax-semantic tagging*. Tesi doctoral. Dep. LSI, UPC.

PALMER, M. (1998). "Consistent criteria for sense distinctions". *Computers and the Humanities* (número especial: *Evaluating Word Sense Disambiguation Programs*). Vol. 1-2, núm. 34, pàg. 217-222.

RESNIK, P.H.; YAROWSKY, D. (1997). "A perspective on word sense disambiguation methods and their evaluation". A: *Workshop: Tagging text with lexical semantics: Why, what and how? (ACL SIGLEX, Washington, D.C.)*.

RIGAU, G.; TAULÉ, M.; GONZALO, J.; FERNÁNDEZ, A. (2001). "Framework and results for the spanish SENSEVAL". A: *Proceedings of the SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems (ACL SIGLEX, Tolosa)*.

TAULÉ, M. (2002). *Especificación de los criterios y la metodología seguida en la organización del SENSEVAL-II español*. X-Tract WP-08/02. Barcelona.

VÉRONIS, J. (1998). "A study of polisemy judgements and inter-annotator agreement". A: *Programme and advanced papers of the SENSEVAL workshop (Herstmonceux Castle, Anglaterra)*.

VÉRONIS, J. (2002). "Sense tagging: does it make sense?". A: *ACL-Workshop on Word Sense Disambiguation (2002: Tolosa)*.

VOSSSEN, P. (ed.). (1999). *EuroEordNet general document* [en línia]. <<http://www.hum.uva.nl/~ewn>>.

Enllaços relacionats:

⇒ SENSEVAL:

<http://www.senseval.org/>

⇒ Grup de Processament del Llenguatge Natural (Universitat de Sheffield):

http://nlp.shef.ac.uk/research/areas/word_sense_disambiguation.html

- ⇒ Article sobre desambiguació semàntica:
<http://www.ilc.cnr.it/EAGLES96/rep2/node39.html>
- ⇒ Demostració de desambiguació semàntica (en alemany):
<http://pi0897.kub.nl/~iris/>

Citació bibliogràfica:

TAULÉ, Mariona; MARTÍ, M. Antònia (2003). "SENSEVAL, una aproximació computacional al significat". *Digitum* [article en línia]. UOC. Núm. 5. [Data de consulta:]

<<http://www.uoc.edu/humfil/cat/art/taule0303/taule0303.html> >

ISSN 1575-2275

[Data de publicació: abril de 2003]