WordNet

ITALICA
Universidad de Sevilla
José A. Troyano

Índice

Introducción

Diccionarios, diccionarios electrónicos y WordNet

Diccionario: Define el léxico de un idioma. Apropiado para uso humano. En la práctica, su uso supone una tarea tediosa.

Diccionario electrónico: Recurso lingüístico. Se puede derivar a partir de los diccionarios tradicionales. Apropiado para el uso automático y humano vía interfaz.

WordNet: Recurso lingüístico ideado para uso automático. Incorpora información psicolingüística. Organizado en base a significados (thesaurus).

Introducción

Lo que aporta WordNet

Diccionarios convencionales:

- Descripciones semánticas (glosa)
- Deletreado
- Pronunciación
- Formas derivadas
- Etimología
- Información gramatical
- Usos
- Sinónimos/antónimos

Lo que aporta WordNet

- Descripciones basadas en conceptos
- Relaciones psicolingüísticas entre palabras

Introducción

Motivación y estructura básica

1985: Un grupo de lingüistas de la universidad de Princeton decidieron construir una base de datos estructurada conforme a criterios psicolingüistas. La idea original era buscar palabras conceptualmente en lugar de alfabéticamente.

WordNet divide el lexicón en cinco categorías:

- nombres
- verbos
- adjetivos
- adverbios
- partículas

Evidentemente hay formas que pueden estar en más de una. Por ejemplo, *close*, puede ser nombre, verbo, adjetivo y adverbio.

Introducción

Semántica léxica

Palabra: Asociación convencional entre un concepto lexicalizado y un lexema (*utterance*) que desempeña un papel sintáctico.

¿qué tipo de lexemas entran dentro de estas asociaciones léxicas? (hay lexemas que no tienen una gran carga de significado, cumpliendo básicamente una función sintáctica)

¿cuál es la naturaleza y organización de los conceptos lexicalizados que pueden expresar las palabras?

¿qué papeles sintácticos juegan las diferentes palabras?

WordNet se centra en la segunda pregunta.

Introducción

Una representación del concepto “palabra”

Parte de la confusión anterior se debe al doble uso del término palabra.

Otra definición de palabra, más matemática, y por tanto más adecuada para una representación formal puede servir para aclarar los conceptos:

palabra = <forma,significado>

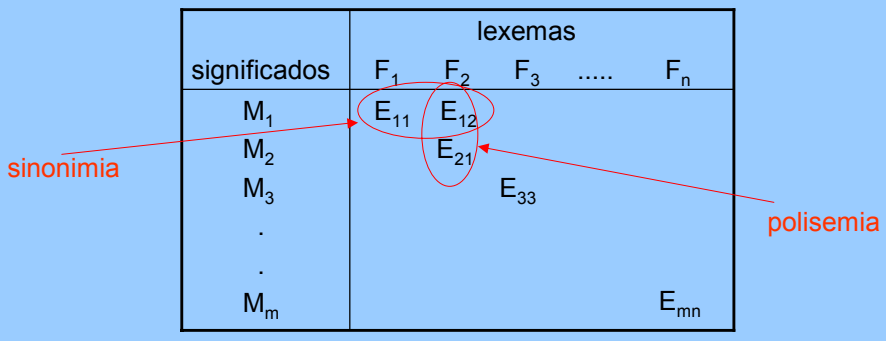
La forma puede ser simple o múltiple (colocación).

De manera que se separa el aspecto de la palabra (*word form*) de lo que significa la palabra (*word meaning*).

Introducción

La matriz léxica

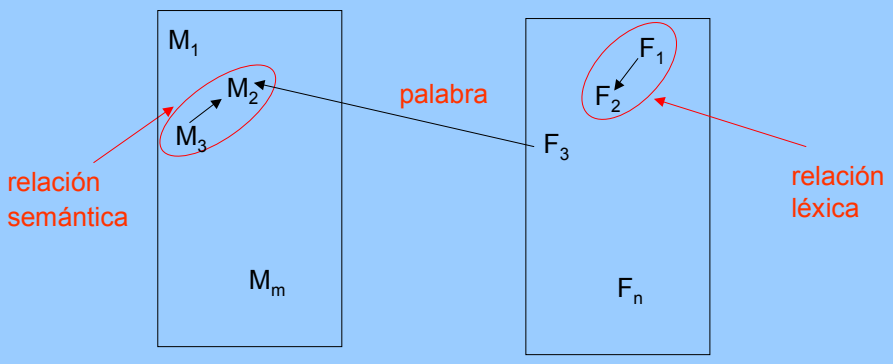
La definición anterior abre las puertas a un sistema de representación que combine las formas y los significados. En esa representación se basa WordNet.



La E es simplemente una entrada y denota la existencia de una relación entre una forma y un significado.

Introducción

Otra posible representación



En un principio WordNet se orientó a la definición de palabras y relaciones semánticas, pero con el tiempo se incorporaron también relaciones léxicas.

Introducción

¿Cómo se representan los significados?

	lexemas			
significados	suelo	piso	territorio	planta
M_1	E_{11}	E_{12}		
M_2	E_{21}		E_{23}	
M_3	E_{31}			
M_4	E_{41}			
M_5		E_{52}		E_{54}

Synsets (synonym sets):

M_1={suelo, piso}

M_2={suelo, territorio}

M_3={suelo, (superficie inferior de algunas cosas; p.e.,de las vasijas.)}

M_5={suelo, (modalidad de gimnasia artística)}

M_4={piso, planta}

Introducción

Relaciones: semánticas, léxicas y morfológicas

WordNet está organizado en base a relaciones. Dado que los significados se representan mediante synsets, las relaciones semánticas se pueden representar mediante enlaces entre synsets.

Las relaciones más importantes contempladas en WordNet son:

- Sinonimia
- Antonimia
- Hiponimia/hiperonimia
- Holonimia/meronimia
- Morfológica

Introducción

Relaciones: sinonimia

Es la relación más importante de WordNet.

Definición: dos expresiones son sinónimas en un contexto C si la sustitución de una por otra en dicho contexto no altera el significado.

La definición de los synsets en términos de sustitución hace necesaria la separación en categorías sintácticas (nombres, verbos, adjetivos y adverbios).

Hay que tener en cuenta que la discretización del concepto sinonimia (dos palabras son sinónimas o no) impide capturar todo el rango de matices que ofrece el lenguaje natural.

Introducción

Relaciones: antonimia

La antonimia se refiere a los contrarios, por ejemplo *rico*/*pobre*.

Al igual que ocurre con la sinonimia, hay que ser cauteloso con la discretización de significados. No siempre la negación de un concepto coincide con su antónimo, por ejemplo no ser *rico* no significa ser *pobre*.

Es una relación entre lexemas, no está claro que lo sea siempre de significados (synsets). Así, con los synsets {blanco,claro} y {negro,oscuro}:

[blanco/negro] son antónimos

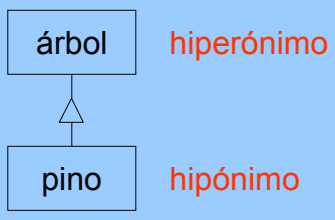
[claro/oscuro] son antónimos

*[blanco/oscuro]

Es muy utilizada en la definición de adjetivos y adverbios.

Introducción

Relaciones: hiponimia/hiperonimia

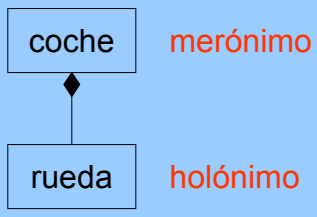


Es una relación transitiva y asimétrica.

Es fundamental en la definición de nombres en WordNet.

Introducción

Relaciones: holonimia/meronimia



Es una relación transitiva y asimétrica.

Se asume que “el concepto de una parte de un todo” puede ser “una parte de un concepto del todo”.

Introducción

Relaciones: morfológica

En el diseño original de WordNet no se contemplaba este tipo de relaciones.

Se incluyeron con idea de hacer práctico y útil el sistema desarrollado.

Por ejemplo si la palabra *árbol* está registrada en WordNet es necesario que ante la palabra *árboles* sea capaz de analizarla morfológicamente para acceder a la información de su forma base.

Introducción

¿Cómo se usa WordNet?

Los lexicógrafos:

- Creando y completando ficheros lexicográficos.

Los usuarios:

- Con una interfaz en C (“wn.h”)
- Con la instrucción wn (en línea de comandos)
- Con una interfaz gráfica.

Índice

Nombres

Estadísticas

En la versión actual (1.7) de WordNet, los números de los nombres son:

Formas distintas:	107930
Synsets:	74488
Parejas forma-significado:	132407
Formas monosémicas:	94025
Formas polisémicas:	13905
Polisemia media:	1’22

Nombres

Lo que no se dice en un diccionario (I)

árbol: planta perenne, de tronco leñoso y elevado, que se ramifica a cierta altura del suelo. (RAE)

no dice:

- que el árbol tiene hojas y raíces.
- que las paredes de sus células están compuestas de celulosa.
- que es un organismo vivo.

Estas cosas se pueden encontrar en el significado de *planta.* Pero ¿cuál de ellos?

planta: (1) Ser orgánico que crece y vive sin mudar de lugar por impulso voluntario. (2) Cada uno de los pisos o altos de un edificio.

Nombres

Lo que no se dice en un diccionario (II)

Tampoco se dice:

- si existen otro tipo de plantas.
- si existen distintos tipos de árboles.
- cuáles son.

Estas cosas se pueden encontrar buscando de la A a la Z otras definiciones que refieran a planta o a árbol.

Tampoco se dice:

- crecen a partir de semillas.
- los ejemplares adultos suelen ser más altos que las personas.
- generan su alimento a partir de la fotosíntesis

Este tipo de cosas se pueden encontrar en enciclopedias o en múltiples fuentes. Aquí un diccionario convencional se queda corto.

Nombres

Lo que no se dice en un diccionario (III)

La mayor parte de la información que no está es estructural, de relaciones entre conceptos.

• Hiperonimia entre planta, árbol y pino.

• Meronimia entre árbol, raíz, hoja y célula.

• Otras relaciones entre árbol y semilla, árbol luz y oxígeno o árbol y persona.

Nombres

Herencia léxica

Las palabras de un idioma están relacionadas entre sí. En muchas ocasiones las definiciones incurren en ciclos.

Mediante la herencia se pueden evitar estructuras circulares y dirigir las definiciones a estructuras arbóreas.

Asunciones psicolingüísticas:

Él tenía una moto, salía a menudo por carretera, pero no tenía licencia para conducir ese vehículo.

Se suelen utilizar hiperónimos como anáforas

¿Canta un canario?
¿Vuela un canario?
¿Tiene piel un canario?

Se tarda más en contestar a preguntas referidas a características del hiperónimo

Nombres

Una red de herencias

La relación de herencia permite definir estructuras arbóreas basándose en la hiperonimia:

olmo @→ árbol @→ vegetal @→ organismo

Esta misma relación se puede ver en sentido inverso (hiponimia):

organismo ~→ vegetal ~→ árbol ~→ organismo

Los synsets de WordNet permiten representar toda una red de relaciones de herencia:

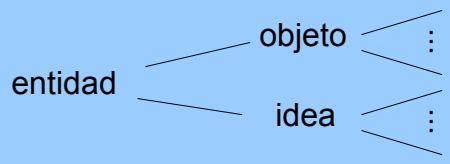
{árbol, planta, @ conífera, ~ abeto, ~...}

{planta, árbol, ~ organismo, @ ...}

Nombres

Componentes semánticos (I)

El recorrido de las relaciones de herencia hacia arriba puede llevarnos, en la raíz, al concepto más general:



Esto puede llevar a conceptos vacíos de significado y clasificaciones antinaturales.

La alternativa de WordNet es la de utilizar un número pequeño de factores primos semánticos, que sean cabecera de jerarquía.

Nombres

Componentes semánticos (II)

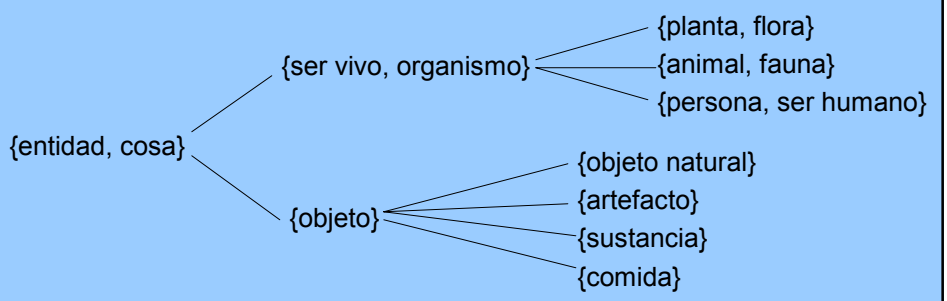
Las 25 raíces de WordNet son (traducidas al español):

{acto, acción, actividad}
{animal, fauna}
{atributo, propiedad}
{cuerpo}
{conocimiento}
{evento, suceso}
{sentimiento, emoción}
{comida}
{grupo, colección}
{lugar}
{motivo}
{artefacto}
{comunicación}
{objeto natural}
{fenómeno natural}
{persona, ser humano}
{planta, flora}
{posesión}
{proceso}
{cantidad}
{relación}
{forma}
{relación}
{estado, condición}
{sustancia}

Nombres

Componentes semánticos (III)

Estas 25 categorías se pueden a su vez clasificar. En WordNet esta clasificación se representa también, aunque el grueso de los nombres se organiza con las 25 categorías originales. Por ejemplo con las cosas:



Nombres

La profundidad de la herencia

Los 25 ficheros resultantes son bastante planos, es raro alcanzar una profundidad mayor de 10.

Las relacionadas con artefactos o cuestiones técnicas suelen ser muy profundas.

En estas jerarquías se suele identificar (más o menos en la mitad) el nivel básico. Más arriba las definiciones son vagas y más abajo demasiado detalladas.

Los nombres situados en este nivel básico se denominan conceptos genéricos.

Nombres

Características

Un sistema que sólo contemple la herencia léxica deja fuera muchos aspectos importantes en la definición de un nombre.

Por ejemplo un canario es un hipónimo de pájaro pero además es:

pequeño, amarillo, canta, vuela, tiene pico y alas

Las características pueden ser de tres tipos:

Atributos (adjetivos): pequeño, amarillo

Partes (nombres): pico, alas

Funciones (verbos): cantar, volar

Nombres

Atributos

Supone una relación entre el nombre y un adjetivo.

No están implementados en WordNet. Denota una relación unidireccional (al menos sólo es útil una de las direcciones):

• Es necesario saber que un canario es amarillo.

• No es tan importante obtener todos los nombres de las cosas amarillas.

En ocasiones, el atributo de un nombre debe ser interpretado sólo con respecto a su inmediato hiperónimo:

Un *canario* es *pequeño* para ser un *pájaro*, pero no es *pequeño* en términos absolutos.

Hay ciertas restricciones de asociación. Tienen bastante que ver con la partición en 25 categorías:

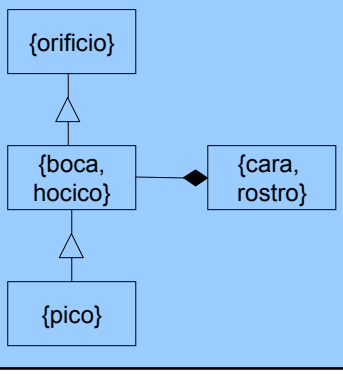
*Este *canario* es *generoso*

Nombres

Partes

La relación parte-de (meronimia/holonimia) es transitiva y asimétrica.

Las partes se heredan de hiperónimo a hipónimo.



En muchas ocasiones, el problema al establecer una adecuada aplicación de meronimia e hiperonimia se debe a la tendencia a asignar características de forma temprana a los conceptos abstractos.

Nombres

Partes: distintas formas de pertenecer a algo

No todas las relaciones parte-de son equivalentes:

- Componente (rama/árbol)
- Miembro de colección (árbol/bosque)
- Material (aluminio/avión)
- Porción (pan/rebanada)
- Actividad (pago/compra)
- Lugar (Sevilla/Andalucía)

Esto hace que la transitividad en ocasiones no esté muy clara o, al menos, “suene mal”.

La rama es parte de un árbol.

El árbol es parte de un bosque.

*La rama es parte de un bosque

Nombres

Partes: relaciones contempladas en WordNet

WordNet sólo implementa tres relaciones parte-de:

W_m #p→ W_h W_h es un componente de W_m

W_m #m→ W_h W_h es un miembro de la colección W_m

W_m #s→ W_h W_h es el material de W_m

La más frecuente de las tres es #p→

Nombres

Partes: ¿cuándo acaba la descomposición?

Si profundizamos en la descomposición podemos llegar al nivel de átomos.

Desde el punto de vista del conocimiento léxico, no está claro que esta finura sirva de ayuda.

La descomposición de un objeto se para cuando las partes no sirven para distinguir al objeto compuesto.

Nombres

Funciones

Supone una relación entre el nombre y un verbo.

No están implementadas en WordNet.

Si interpretamos esta acción del nombre como la función que lo caracteriza hay distintas situaciones:

La función del lápiz es escribir.

La función de un canario es cantar. ¿?

La función de un adorno es decorar. (Da más información que cualquier otra propiedad de adorno).

Hay que evitar la circularidad, sobretodo en inglés (*butter* verbo y nombre).

Nombres

Antonimia

Los psicolingüistas determinan que una palabra es antónima de otra si es la respuesta más frecuente en un test de asociación.

No es una relación importante a nivel semántico pero no es difícil integrarla en un esquema como el de WordNet:

{[hombre,mujer!], persona, @...}

{[mujer,hombre!], persona, @...}

Esta relación se hereda a los correspondientes hipónimos, creando nuevas parejas de antónimos:

padre/madre, marido/mujer, rey/reina,...

Índice

- Introducción
- Nombres
- **Adjetivos**
- Verbos
- Diseño e implementación

Adjetivos

Modificadores de nombres

La función principal de un adjetivo es la de modificar el significado de un nombre.

Otras categorías sintácticas también pueden desempeñar la función del adjetivo. Se da más en inglés que en español:

Inglés	Español
creaking chair	silla chirriante
painted chair	silla pintada
barber chair	silla de barbero
chair by the window	silla junto a la ventana
my grandfather's chair	la silla de mi abuelo

Los synsets de adjetivos de WordNet contienen básicamente adjetivos, más algunos nombres y frases preposicionales.

En total hay unas 21000 formas distintas agrupadas en unos 18000 synsets.

Adjetivos

Adjetivos descriptivos

Son aquellos que asocian un valor a un atributo de un nombre. Es decir, si el nombre "*N* es *adj*" entonces hay un atributo *A* tal que:

$$A(N)=adj$$

Por ejemplo decir "el cajón es pesado" presupone que *cajón* tiene un atributo *PESO* cuyo valor es *pesado*.

WordNet implementa esta relación mediante punteros que relacionan los synsets de los nombres de los atributos y los adjetivos que los describen.

Adjetivos

Adjetivos descriptivos: hiponimia vs antonimia

En los adjetivos la relación hiponimia/hipernonimia es menos natural que en los nombres.

No está claro que un adjetivo "sea un subtipo" de otro adjetivo.

En el caso de los adjetivos descriptivos la antonimia sí se presenta como una relación interesante:

bueno - malo
ligero - pesado

La antonimia representa la bipolaridad de los atributos. En WordNet esta oposición se representa con la relación !→

bueno !→ malo
malo !→ bueno

Adjetivos

Adjetivos descriptivos: hay que matizar la antonimia

-¿Por qué adjetivos de significado muy cercano tienen antónimos distintos?

heavy (pesado), weighty (de peso) tienen como antónimos a light (ligero) y weightless (ingrávido) respectivamente.

-¿Por qué hay muchos adjetivos descriptivos que no tienen antónimos?

ponderous(laborioso, lento, pesado) podría tener como antónimo a light, pero light ya tiene su antónimo.

Esto induce a pensar que hay otro tipo de relación involucrada en la definición de los adjetivos.

Adjetivos

Adjetivos descriptivos: antonimia entre palabras

Las anteriores preguntas impiden establecer relaciones de antonimia entre synsets. Así ante los synsets:

{heavy,weighty,ponderous} {light,weightless,airy}

Se admitirían las parejas de antónimos:

weigthy/weightless, heavy/light

Pero serían difícilmente admisibles otras como:

heavy/weightless, ponderous/airy(etéreo)

Para solucionar esto, WordNet organiza los adjetivos en synsets (por similitud), mientras que para reflejar una relación de la antonimia se elige una palabra que “representa” al synset.

Adjetivos

Adjetivos descriptivos: antonimia directa e indirecta

El planteamiento anterior lleva a dos tipos de antonimia entre adjetivos:

-directa: la que se establece entre dos palabras

-indirecta: la que se hereda por pertenecer a un synset (por similitud)

La relación de antonimia directa se expresa mediante !→. Por su parte la relación de similitud se expresa con &→. De esta forma expresaríamos que moist(húmedo) es antónimo indirecto de dry(seco):

moist&→wet!→dry

Mientras que wet(mojado) es antónimo directo de dry:

moist!→wet!→dry

Adjetivos

Adjetivos descriptivos: gradación

Adjetivos contradictorios: no pueden ser ciertos a la vez, pero tampoco falsos a la vez (vivo/muerto)

Adjetivos contrarios: no pueden ser ciertos a la vez, pero sí falsos a la vez (gordo/flaco).

El problema de la definición de contrario es que es muy laxa, y no se limita a los opuestos. Por ejemplo *gaseoso* y *vegetal* serían contrarios.

Para aclarar estas ideas se hace necesario el concepto de gradación o escala de un atributo.

Adjetivos contradictorios: Están en la misma escala y no son graduables.

Adjetivos contrarios: Están en la misma escala y son graduables.

Adjetivos

Adjetivos descriptivos: escalas

tamaño	edad	temperatura
astronomical	ancient	torrid
huge	old	hot
large	middle-aged	warm
standard	mature	tepid
small	adolescent	cool
tiny	young	cold
infinitesimal	infantile	frigid

Se calcula que no más del 2% de los adjetivos de WordNet se puede graduar.

Lo normal es utilizar un adverbio para graduar el adjetivo.

WordNet no implementa la gradación.

Adjetivos

Adjetivos descriptivos: marcado

Muchos atributos tienen asociada algún tipo de dimensión. En una pareja de antónimos hay cierta asimetría al respecto:

The road is ten miles long

*The road is ten miles short

long está asociado a la dimensión (incluso morfológicamente con *lenght*) y *short* no. Se dice que *long* está marcado.

Esta relación es obvia cuando se utilizan prefijos negativos:

un+pleasant im+patient il+legal

WordNet no implementa el marcado

Adjetivos

Adjetivos descriptivos: calificación selectiva

No todos los atributos pueden calificar a todos los nombres:

un relato alto, ¿un relato corto?

Hay adjetivos muy generales:

bueno,malo

de gran ámbito de aplicación:

activo, pasivo

y muy específicos:

atornillado, deshilachado

A pesar de que WordNet tiene una categorización de nombres, no se aprovecha para modelar este concepto.

Adjetivos

Limitaciones sintácticas

Los adjetivos descriptivos suelen ser sintácticamente libres. Pueden ser utilizados de forma atributiva (o prenominal):

big house

O de forma predicativa:

This house is big

Esto no ocurre con otros tipos de adjetivos:

-modificadores de referencia

-relacionales

que se utilizan mayormente de forma atributiva.

Adjetivos

Adjetivos modificadores de referencia (I)

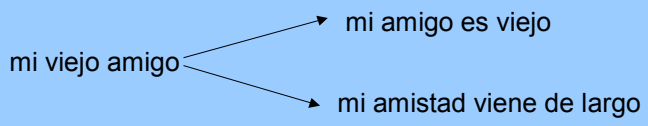
Referente:Ser u objeto de la realidad extralingüística a los que remite el signo.

Referencia:Combinación de signos que identifican un objeto.

En algunas situaciones el adjetivo sólo modifica a la referencia y no al referente:

El anterior presidente

Puede dar lugar a ambigüedad:



Adjetivos

Adjetivos modificadores de referencia (II)

-Este tipo de adjetivos no es muy numeroso (unas pocas docenas)

- Suelen referirse a estados temporales de los nombres.

- O denotan algún tipo de conocimiento (supuesto, potencial)

- Pueden desempeñar la función de un adverbio:

mi antiguo profesor => fue antiguamente mi profesor

- En WordNet los adjetivos modificadores de referencia están marcados como no predicativos.

Adjetivos

Adjetivos de color

Son un tipo muy particular de adjetivos:

- Pueden ser nombres y adjetivos a la vez.

- Pueden ser graduados.

- Pueden combinarse con otros adjetivos descriptivos.

- No están sujetos a antonimia (ni directa ni indirecta), excepto en algunos atributos como *claro/oscuro*, *brillo/mate*.

Adjetivos

Adjetivos relacionales (I)

- Son adjetivos relacionados o pertinentes a una determinada cosa o concepto. Por ejemplo fraternal se refiere a un hermano o dental ser refiere a diente.

- Se utilizan mayormente de forma atributiva.

- Algunas veces un mismo adjetivo puede ser utilizado como descriptivo y como relacional:

  ley criminal

  comportamiento criminal, él es un criminal

- En inglés los adjetivos relacionales suelen derivar del griego y del latín.

Adjetivos

Adjetivos relacionales (II)

- A diferencia de los descriptivos, no están asociados a un atributo.
- No tienen antónimos directos (aunque pueden formarse con la partícula *non*-).
- No son graduables.
- WordNet mantiene un fichero separado para adjetivos relacionales asociándolos a sus correspondientes nombres.
- Existen unos 1700 synsets que agrupan unas 3000 formas.
- Pueden ser utilizados de forma predicativa en contextos comparativos:

  Estas armas no son químicas ni biológicas, son nucleares.

Adjetivos

Codificación: adjetivos descriptivos (I)

- Se organizan en clusters bipolares, con un polo para cada antónimo.
- Cada polo del cluster está encabezado por una cabecera, que contiene las palabras clave en mayúsculas seguidas de otros adjetivos relacionados por similitud (&).
- Después de la cabecera se incluyen distintos synsets que completan el significado del cluster.
- Se utilizan números para diferenciar distintos sub-significados de una determinada palabra o forma.

Adjetivos

Codificación: adjetivos descriptivos (II)

```
[{ [WET1, DRY1,!] bedewed,& boggy,& clammy,& damp,& drenched,&
drizzling,& hydrated,& muggy,& perspiring,& saturated2,&
showery,& tacky,& tearful,& watery2,& WET2,& }
{ bedewed, dewy, wet1,& }
{ boggy, marshy, miry, mucky, muddy, quaggy, swampy, wet1,& }
{ clammy, dank, humid1, wet1,& }
{ damp, moist, wet1,& }
{ drenched, saturated1, soaked, soaking, soppy, soused, wet1,&
}
{ drizzling, drizzly, misting, misty, wet1,& }
{ hydrated, hydrous, wet1,& ((chem) combined with water
molecules) }
{ muggy, humid2, steamy, sticky1, sultry, wet1,& }
{ perspiring, sweaty, wet1,& }
{ saturated2, sodden, soggy, waterlogged, wet1,& }
{ showery, rainy, wet1,& }
{ sticky2, tacky, undried, wet1,& ("wet varnish") }
{ tearful, teary, watery1, wet1,& }
{ watery2, wet1,& (filled with water; "watery soil") }
```

Adjetivos

Codificación : adjetivos descriptivos (III)

```
-{
[DRY1, WET1,!] anhydrous,& arid,& dehydrated,& dried,& dried-
up1,&
dried-up2,& DRY2,& rainless,& thirsty,& }
{ anhydrous, dry1,& ((chem) with all water removed) }
{ arid, waterless, dry1,& }
{ dehydrated, desiccated, parched, dry1,& }
{ dried, dry1,& ("the ink is dry") }
{ dried-up1, dry1,& ("a dry water hole") }
{ dried-up2, sere, shriveled, withered, wizened, dry1,&
(used of vegetation) }
{ rainless, dry1,& }
{ thirsty, dry1,& }]
```

Adjetivos

Codificación: restricciones y enlaces con clusters (I)

Restricciones sintácticas

- No se expresan para los synsets sino para las formas individuales.

- Los adjetivos que sólo se pueden usar de forma predicativa se marcan con (p).

- Los adjetivos que sólo se pueden usar de forma atributiva se marcan con (a).

- Los pocos adjetivos que (en inglés) pueden utilizarse de forma posnominal se marcan con (ip).

Vínculos con otros clusters:

- Se incluyen en la cabecera.

- Se interpretan como “ver también...”.

Adjetivos

Codificación : restricciones y enlaces con clusters (II)

```
[{ [AWAKE(p), ASLEEP,!] ALERT,& astir(p),& AWARE(p),& CONSCIOUS,&
insomniac,& unsleeping,& }
{ astir(p), out_of_bed(p), up(p), awake,& }
{ insomniac, sleepless, wakeful, awake,& }
{ unsleeping, wide-awake, awake,& }
-{
[ASLEEP(p), AWAKE,!] at_rest(p),& benumbed,& DEAD,& dormant,&
drowsing,&
drowsy,& unconscious,& UNAWARE,& UNCONSCIOUS,& }
{ at_rest(p), resting, asleep,& }
{ benumbed, insensible, numb, unfeeling, asleep,& ("my foot is
asleep") }
{ dormant, inactive, hibernating, torpid, asleep,& }
{ drowsing, dozing, napping, asleep,& }
{ drowsy, nodding, sleepy, slumberous, slumbrous, somnolent,
asleep,& }
{ unconscious, asleep,& }]
```

Índice

- Introducción
- Nombres
- Adjetivos
- **Verbos**
- Diseño e implementación

Verbos

Características básicas

- Probablemente es la categoría sintáctica más importante de los lenguajes.

- Muchos lingüistas apuestan por un modelo semántico en el que el verbo ocupa el lugar central.

- El verbo relaciona al resto de los elementos de la frase.

- Esta papel tan importante debe recogerse en la información léxica asociada al verbo.

- No es una categoría muy numerosa WordNet contiene unos 10000 verbos agrupados en unos 12.000 synsets (la mitad que de adjetivos y una décima parte que de nombres).

- Son muy polisémicos, 2’15 significados/forma (1’22 en nombres, 1’45 en adjetivos, 1’24 en adverbios).

Verbos

Polisemia

- Los significados de los verbos son más flexibles que los de otras categorías.

- Si una persona parafrasea una frase es más fácil que cambie el verbo que los nombres.

- El significado del verbo suele depender de los nombres que lo acompañan.

- Los verbos más usados (have, be, run, make, set, go, take, get,...) son también los más polisémicos:

I have a Mercedes
I have a headache
→ La polisema la marca más la naturaleza del objeto que el verbo

- Para reducir la ambigüedad sería deseable vincular los verbos con los nombres a los que se aplica. WordNet no lo hace.

Verbos

Organización

- Se organizan en 15 ficheros (categorías) según un criterio semántico:

cuidado corporal y fisiología	emoción
cambio	movimiento
conocimiento	percepción
comunicación	posesión
competición	relaciones sociales
consumo	meteorología
contacto	estados (parecerse, bastarse
creación	pertenecer,...)

Los ficheros contienen pequeños clusters semánticos (al estilo de los adjetivos).

Dentro de cada fichero se diferencian verbos de evento y estado.

Verbos

Sinonimia

- Hay muy pocos verbos realmente sinónimos como *close/shut*.
- En inglés, los casos más comunes suelen ser parejas de procedencia Grecolatina y anglosajona: *begin/commence*.
- Las formas grecolatinas suelen ser más formales.
- Los cambios de matiz suelen obedecer a restricciones de selección:

  rise the temperature/ *ascend the temperature

- Debido a todo esto los synsets de los verbos a menudo expresiones perifrásicas en lugar de sinónimos.
- Estas expresiones aportan información léxica introduciendo palabras relacionadas con el significado del verbo:

  {hammer, hit with a hammer}

  {swim, travel through water}

Verbos

Análisis semántico

<u>Compositivo</u>

- Es un modelo generativo. Asume la existencia de unas acciones primas (primitivas) universales.
- Se define el significado de un verbo como combinación de esas acciones primas.

  kill : CAUSE TO BECOME NOT ALIVE

- Tiene bastantes detractores.

<u>Relacional</u>

-Considera las propias formas léxicas como las unidades básicas de significado.

- Establece el significado en base a relaciones de similitud.
- WordNet adopta esta postura aunque introduce algunos aspectos compositivos en forma de relaciones semánticas.

Verbos

Vinculación léxica (I)

- Es la vía que usa WordNet para organizar los verbos (al igual que la hiponimia en los nombres y la oposición en los adjetivos).
- Ayuda a incorporar aspectos compositivos en el esquema relacional.
- La vinculación léxica es a los verbos lo que la implicación a los predicados:

  $P \Rightarrow Q$ (es imposible P cierto y Q falso)

  $V_1 * V_2$ (es imposible hacer V_1 y no hacer V_2)

  roncar * dormir
- La vinculación léxica es unilateral. Excepto en el caso de verbos sinónimos que es bilateral (doble implicación).
- La negación cambia el sentido de la vinculación.
- Recuerda un poco al papel de la meronimia en los nombres.
- Indica una subactividad dentro de otra actividad.

Verbos

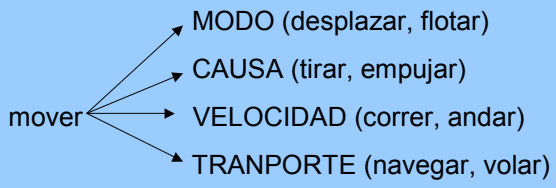
Hiponimia

Parece más natural en los nombres:

un perro es un animal

pasear es andar (parece faltar *despacio* para completar la frase)

De alguna forma, la hiponimia en los verbos está acompañada de otras relaciones semánticas (adverbiales).



A este tipo de hiponimia se le denomina troponimia.

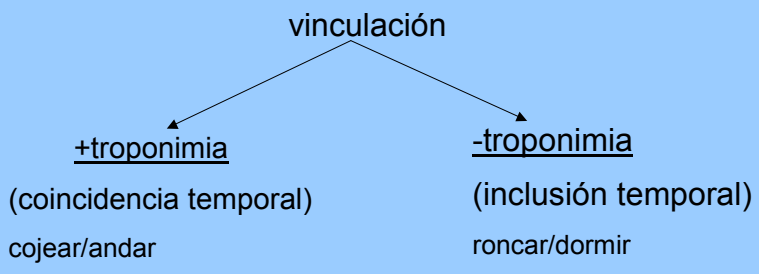
Verbos

Troponimia

V_1 es un tropónimo de V_2 si

V_1 es V_2 de una manera particular

- La troponimia e un tipo particular de vinculación léxica. Si V_1 es un tropónimo de V_2, entonces V_1 implica V_2.



Verbos

Taxonomías de verbos

-No es tan simple construir una taxonomía como la de los nombres.

- Para cada campo semántico no puede encontrarse una única raíz para todos los verbos. Por ejemplo para movimiento y posesión tendríamos dos y tres raíces respectivamente:

{mover, realizar un movimiento} {mover, viajar}

{dar,transferir} {tomar, recibir} {tener, poseer}

- La flexibilidad de los significados de los verbos hace que la jerarquía sea más borrosa que la de los nombres.

- En cualquier caso, en cada rama de la jerarquía es posible encontrar un nodo más lexicalizado, similar a lo que se denomina nivel básico para los nombres.

Verbos

Verbos opuestos

- Muchos de los opuestos se forman con prefijos:

  aparecer/desaparecer atar/desatar

- En inglés, los verbos formados a partir de adjetivos con los sufijos -ify y –en heredan sus opuestos:

  lengthen/shorten prettify/uglify

- Algunos verbos opuestos comparten la misma vinculación léxica:

  fallar implica apuntar

  acertar implica apuntar

- A este tipo de vinculación se le denomina de “condición anterior” (*backward presupposition*).

Verbos

Relación causal

- Esta relación vincula a una pareja de verbos:
  - •causativo: provocan un cambio de estado, p.e. *dar*
  - •resultante: denotan el efecto del cambio, p.e. *tener*
- A diferencia de otras relaciones de WordNet el sujeto del verbo causativo no siempre coincide con el del resultante.
- La relación causal es un tipo especial de vinculación léxica. Si V_1 es causa de V_2, entonces V_1 implica V_2.

  dar implica tener (teniendo en cuenta el cambio del sujeto)

Verbos

Propiedades sintácticas

- WordNet incluye para cada synset patrones que describen las restricciones sintácticas del correspondiente verbo:

  somebody ___ something Adjective/Noun

  somebody ___ somebody with something

- Esta información permite busca los verbos según propiedades sintácticas

- Este tipo de información es útil para el análisis sintáctico y la interpretación semántica.

Índice

Diseño e implementación

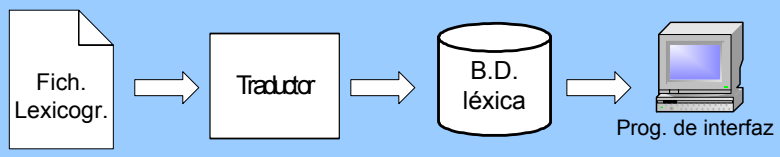
Arquitectura

WordNet está pensado para modelar el conocimiento léxico que puede tener un hablante nativo del inglés.

Su desarrollo se ha dividido en dos tareas:

-Escritura de los ficheros que contienen toda la información léxica.

-Creación de software que interpreten dichos ficheros y ofrezcan servicios a los usuarios.



Diseño e implementación

Palabras más familiares (I)

- Algunas palabras son mucho más familiares que otras.

- Aprovechar bien esta información influye sustancialmente en el rendimiento del sistema.

- WordNet asocia un índice de “familiaridad” a cada palabra.

- En principio parece que esta información se puede extraer de las frecuencias de uso en un corpus.

- Gracias a la ley de Zipf (cuanto más frecuente es una palabra más polisémica es). Se puede también utilizar la polisemia como índice de familiaridad.

Diseño e implementación

Palabras más familiares (II)

Palabra	Polisemia
bronco	1
@→ mustang	1
@→ pony	5
@→ horse	14
@→ equine	0
@→ odd-toed ungulate	0
@→ placental mamma	1 0
@→ mammal	1
@→ vertebrate	1
@→ chordate	1
@→ animal	4
@→ organism	2
@→ entity	3

Diseño e implementación

Ficheros lexicográficos (I)

- Están escritos por lexicógrafos.
- Recogen toda la información léxica.
- Hay dos elementos básicos: significados (syntes) y palabras (word forms).
- Hay dos tipos básicos de relaciones: léxicas (entre palabras) y semánticas (entre synsets).
- Los adverbios están en un único fichero.
- Los verbos y los nombres están organizados en distintos ficheros según criterios semánticos.
- Los adjetivos están organizados en dos ficheros: descriptivos y relacionales.

Diseño e implementación

Ficheros lexicográficos (II)

`noun.Tops`	unique beginners for nouns
`noun.act`	nouns denoting acts or actions
`noun.animal`	nouns denoting animals
`noun.artifact`	nouns denoting man-made objects
...	
`verb.cognition`	verbs of thinking, judging, analyzing, etc.
`verb.communication`	verbs of telling, asking, ordering, singing, etc.
`verb.competition`	verbs of fighting, athletic activities, etc.
`verb.consumption`	verbs of eating and drinking
...	
`adj.all`	all adjective clusters
`adj.pert`	relational adjectives (pertainyms)
`adv.all`	all adverbs

Diseño e implementación

Ficheros lexicográficos (III)

- Cada uno de ellos está compuesto de una serie de synsets.
- De manera general los synsets incluyen sinónimos, glosas y punteros relacionales (hiponimia, hiperonimia,...).
- Los adjetivos descriptivos están organizados en clusters, que representan los valores de un determinado atributo.
- Los clusters contienen dos (en algunos casos tres) partes, etiquetadas con una cabecera identificativa (par de antónimos).
- La aplicación que interpreta estos ficheros y los compila en una base de datos se llama *grind* (rollo, paliza).
- Se distribuye gratuitamente la base de datos y la interfaz, pero no los ficheros lexicográficos ni la aplicación *grind*.

Diseño e implementación

Palabras

-Representación ortográfica de una palabra individual (*awake*).

- Representación ortográfica de colocaciones (dos o más palabras) separadas por subrayados (*out_of_bed*).

- Algunas veces se añade un número al final de una palabra para diferenciar significados.

- A los adjetivos se les puede añadir una marca sintáctica entre paréntesis (a), (p) ó (ip).

Diseño e implementación

Punteros relacionales

- Representan las relaciones entre palabras y significados.

- Las relaciones léxicas entre distintas categorías son:
  - Adjetivos relacionales y sus correspondientes nombres
  - Adverbios y adjetivos de los que derivan

-Las relaciones semánticas entre distintas categorías son:
  - Adjetivos y nombres de atributos
  - Nombres de atributos y adjetivos

-El resto de relaciones se establecen dentro de la misma categoría sintáctica:
  - Se aprovecha el synset para la sinonimia
  - El resto se modela también dentro del synstet con la ayuda de operadores

Diseño e implementación

Gramática simplificada del synset

synset : ‘{‘ elementos ‘}’

elementos : elemento

| elementos elemento

elemento: PALABRA ‘,’

| relacion_semantica

| relacion_lexica

relacion_semantica : PALABRA ‘,’ OPERADOR

relacion_lexica : ‘[‘ PALABRA ‘,’ OPERADOR ‘]’ ‘,’

Diseño e implementación

Operadores de relación

Noun	Verb	Adjective	Adverb
Antonym !	Antonym !	Antonym !	Antonym !
Hyponym ~	Troponym ~	Similar &	Derived from \
Hypernym @	Hypernym @	Relational Adj. \	
Meronym #	Entailment *	Also See ^	
Holonym %	Cause >	Attribute =	
Attribute =	Also See ^		

Diseño e implementación

Relaciones recíprocas

Son generadas automáticamente en la base de datos aunque sólo esté codificado un sentido en los ficheros lexicográficos

Relación	Recíproca
Antonym	Antonym
Hyponym	Hypernym
Hypernym	Hyponym
Holonym	Meronym
Meronym	Holonym
Similar to	Similar to

Diseño e implementación

Sistema de archivos

El sistema de archivos lexicográficos está basado en RCS (Unix Revision Control System). Permite:

- Llevar una historia de modificaciones
- Reconstruir cualquier versión anterior de WordNet
- Prevenir conflictos de escritura entre distintos lexicógrafos

Constan de una serie de scripts Unix que sirven de interfaz al usuario. Por ejemplo:

reserve: extrae la versión más reciente de un archivo y lo bloquea.

review: extrae la versión más reciente de un archivo para consultarlo y no lo bloquea.

Diseño e implementación

La aplicación *grind*

- Compila los ficheros lexicográficos y genera la base de datos léxica.

- Está escrita en C, lex y yacc.

- Identifica errores sintácticos y estructurales (semántica estática).

- Calcula el índice de familiaridad (polisemia).

- La representación interna (sintaxis abstracta) está basada en una tabla hash de palabras.

Diseño e implementación

La base de datos de WordNet

-Está compuesta de ficheros ASCII legibles tanto para personas como para máquinas.

- Consta de ocho ficheros:

index.noun	data.noun
index.verb	data.verb
index.adj	data.adj
index.adv	data.adv

-Cada fichero índice es una lista ordenada alfabéticamente de todas las palabras de una categoría sintáctica.

- Los ficheros de datos contiene la información lexicográfica asociada.

Diseño e implementación

Los ficheros índice

- Palabra, índice de polisemia, relaciones en las qué está inmersa, claves.

...

bipolar_disorder n 1 2 @ ~ 1 0 10327371

biprism n 1 2 @ %p 1 0 02291181

biquadrate n 1 1 @ 1 0 09886612

biquadratic n 3 1 @ 3 0 09886612 05001973 04511971

biquadratic_equation n 1 1 @ 1 0 05001973

biquadratic_polynomial n 1 1 @ 1 0 04511971

birch n 3 5 @ ~ #m #s %s 3 0 08585960 08585601 02291288

...

Diseño e implementación

Los ficheros de datos

- Clave, punteros relacionales, glosas, patrones verbales:

...

00034867 04 n 01 surfacing 0 001 @ 00034703 n 0000 | emerging to the surface and becoming apparent

00034968 04 n 03 dispatch 0 despatch 0 shipment 0 002 @ 00029100 n 0000 ~ 00035095 n 0000 | the act of sending off something

00035095 04 n 01 reshipment 0 001 @ 00034968 n 0000 | the act of shipping again (especially by transferring to another ship)

00035222 04 n 01 completion 1 002 @ 00020977 n 0000 ~ 00035376 n 0000 | the act of becoming or making complete: "her work is still far from completion"

...

Diseño e implementación

¿Cómo se ejecuta WordNet?

wn word [-hgla] [-n#] -searchtype [-searchtype...]

-h	Display help text before search output
-g	Display gloss
-l	Display license and copyright notice
-a	Display lexicographer file information
-o	Display synset offset
-s	Display sense numbers in synsets
-n#	Search only sense number #

searchtype is at least one of the following:

-ants{n|v|a|r}	Antonyms
-hype{n|v}	Hypernyms
-hypo{n|v}, -tree{n|v}	Hyponyms & Hyponym Tree
...	

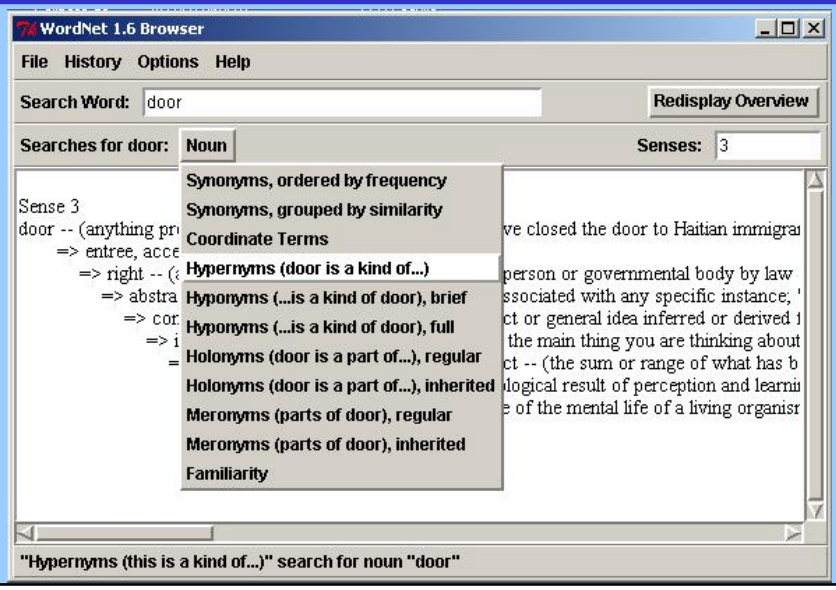
Diseño e implementación

La interfaz: wn.h

```
/* Primary search algorithm for use with user interfaces */

extern char *findtheinfo(char *, int, int, int);

/* Primary search algorithm for use with programs (returns data structure) */

extern SynsetPtr findtheinfo_ds(char *, int, int, int);

/* Set bit for each search type that is valid for the search word   passed and
return bit mask. */

extern unsigned long is_defined(char *, int);

/* Set bit for each POS that search word is in.  0 returned if   word is not in
WordNet. */

extern unsigned int in_wn(char *, int);

...
```

Diseño e implementación

La interfaz gráfica



Diseño e implementación

Análisis morfológico: morphy

Noun		Verb		Adjective	
Sufffix	Ending	Suffix	Ending	Suffix	Ending
s		s		er	
ses	s	ies	y		est
xes	x	es	e	er	e
zes	z	es		est	e
ches	ch	ed	e		
shes	sh	ed			
		ing	e		
		ing			

- También se usa una lista de excepciones para cada categoría gramatical (excepto los adverbios).