Extracción de Información

- Extracción de entidades con nombre
- Instanciación de Entidades con Nombre
- Extracción de relaciones

Instanciación de entidades con nombre

- Uruguay
  - País
  - Río
  - Calle
  - Nombre de persona

- Los nombres propios no son identificadores unívocos de entidades del mundo.

Instanciación de entidades con nombre

- Una vez que hemos reconocido un nombre de entidad, debemos **desambiguarlo**.
- La **clasificación** (Persona, Lugar Geográfico, Organización, etc.) es un 1er paso en la desambiguación, pero no basta.
- Para desambiguar, necesitamos un **repertorio de valores posibles**. En el caso de palabras comunes, estos repertorios son los **diccionarios**.
- Se ha usado la wikipedia como fuente de conocimiento para la desambiguación : ***wikification***

Wikification

- Vincular entidades con nombre en textos con referentes en la wikipedia.
- No todas los referentes van a estar descriptos en la wikipedia, pero si los que involucren conocimiento “enciclopédico”.
- Otras desambiguaciones se harán en un contexto local

Wikification

- Wikipedia, “Montevideo desambiguación”

Por **Montevideo** pueden entenderse los siguientes conceptos:

Geográficos [editar]

- **Montevideo**, ciudad capital del Uruguay.
- **Departamento de Montevideo**, departamento que incluye a la ciudad anterior.
- **Área Metropolitana de Montevideo**, conglomerado urbano que incluye localidades cercanas.
- **Cerro de Montevideo**, cerro de la ciudad homónima.
- **Cerro Montevideo**, cerro de las Islas Malvinas.
- **Bahía de Montevideo**.
- **Montevideo (Estados Unidos)**, localidad de Minnesota en Estados Unidos.
- **Distrito de Montevideo**, en la provincia de Chachapoyas, Departamento de Amazonas, Perú.
- **Montevideo Chico**, localidad uruguaya del departamento de Tacuarembó.

Extracción de relaciones

- Identificadas las entidades se busca extraer relaciones entre estas.
- La mayor parte de los trabajos:
  - relaciones entre entidades mencionadas en la misma oración
  - relaciones predeterminadas (dirección de una empresa, empresa donde trabaja una persona, etc.)
  - relaciones binarias; se habla de extracción de eventos cuando hay más de 2 argumentos

Extracción de relaciones

El presidente **Tabaré Vázquez** se reunirá hoy con su par brasileño **Michel Temer** en **Nueva York**, a donde viajó para participar de la **Asamblea de Naciones Unidas**. También se reunirá con la directora general de la **OMS**, **Margaret Chan.**

1) Relación : reuniones de mandatarios o personalidades
2) 2 argumentos: A se reúne con B
3) En la última oración no funcionaría a no ser que se recupere el sujeto omitido.
4) Podríamos agregar DONDE y CUANDO es la reunión. Ya no sería extracción de relaciones sino de **eventos**

Extracción de relaciones

El presidente **Tabaré Vázquez** se reunirá hoy con su par brasileño **Michel Temer** en **Nueva York**, a donde viajó para participar de la **Asamblea de Naciones Unidas**. También se reunirá con la directora general de la **OMS**, **Margaret Chan**.

- Otras relaciones que se informan:
    - presidente de
    - director general de
    - viaje a

Conferencia ACE 2008

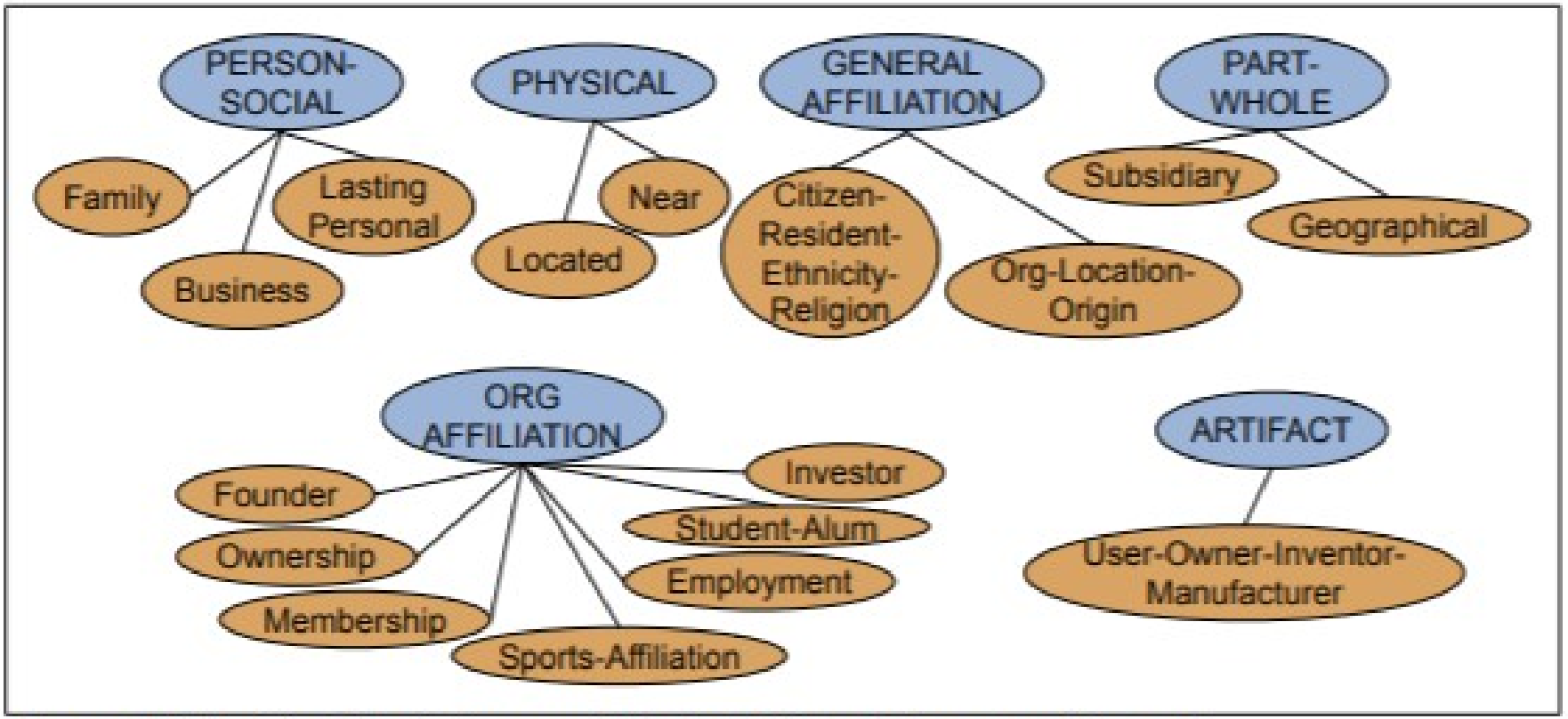


Figure 18.9 The 17 relations used in the ACE relation extraction task.

Relaciones agrupadas en clases : p.ej., hay 7 relaciones en la clase afiliación a una organización.

Conferencia ACE 2008

Relations	Types	Examples
Physical-Located	PER-GPE	**He** was in **Tennessee**
Part-Whole-Subsidiary	ORG-ORG	**XYZ**, the parent company of **ABC**
Person-Social-Family	PER-PER	**Yoko**'s husband **John**
Org-AFF-Founder	PER-ORG	**Steve Jobs**, co-founder of **Apple**...

Figure 18.10 Semantic relations with examples and the named entity types they involve.

Notar que las relaciones que se extraen son proposiciones contingentes, o sea, predicados aplicados a argumentos que se hacen verdaderos o falsos al instanciarse en una situación y entidades específicas.

En extracción de relaciones no se trata el tema de los intervalos temporales de validez de una relación, pero si cuando se trata de extraer eventos o estados.

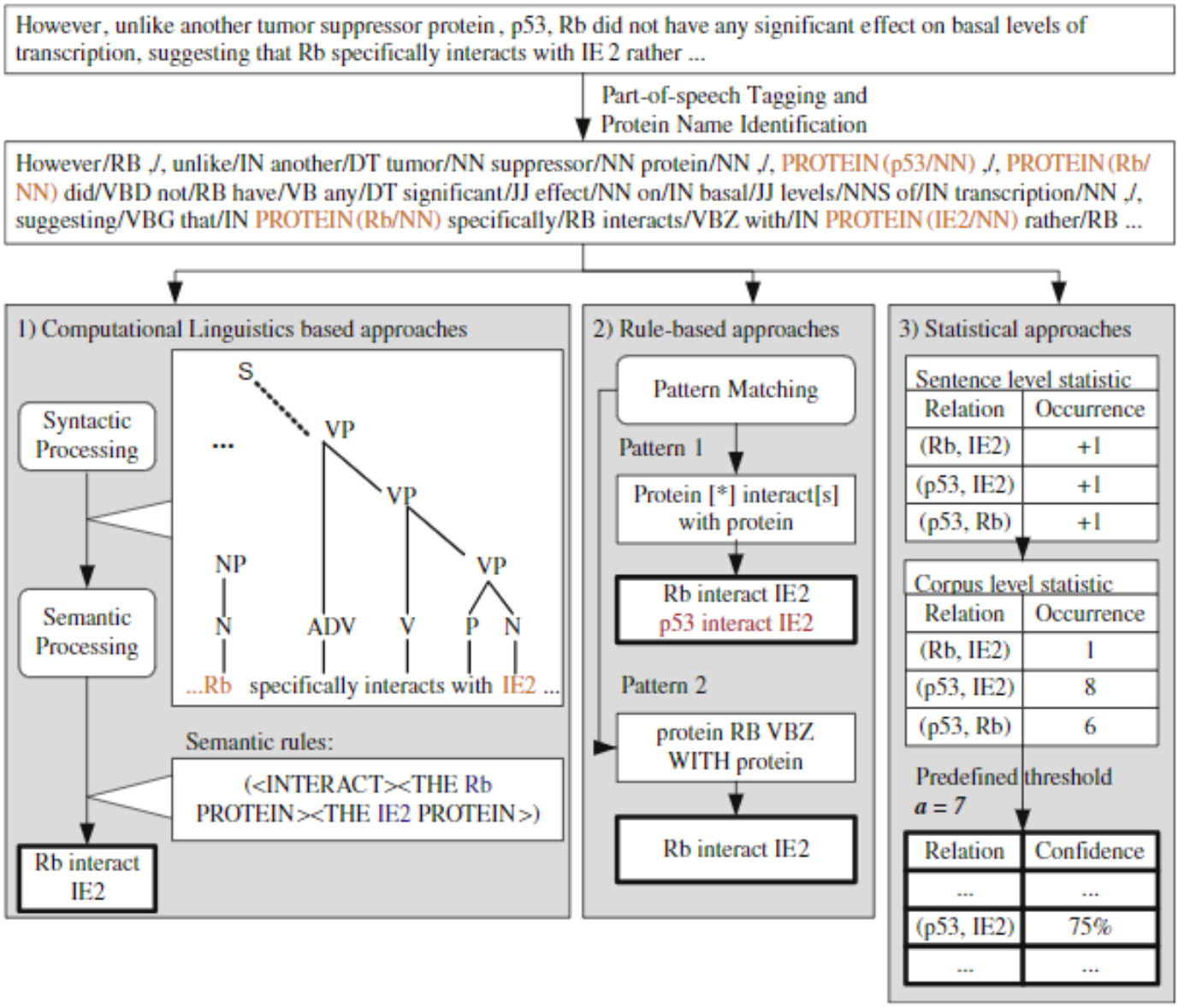
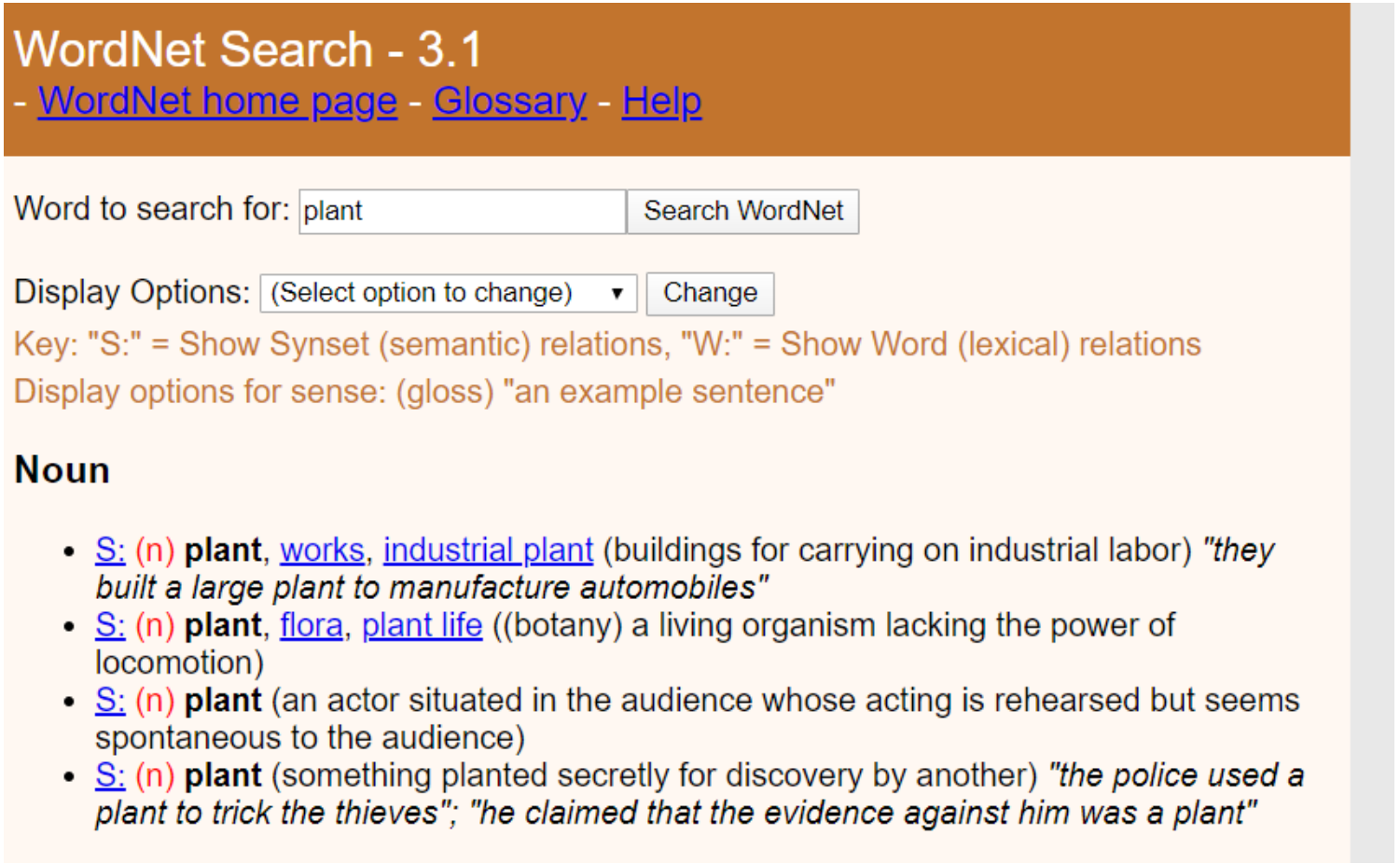


Fig. 3. General dataflow of information extraction system employing different methodologies.

WordNet, acepciones



WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: plant Search WordNet

Display Options: (Select option to change) Change

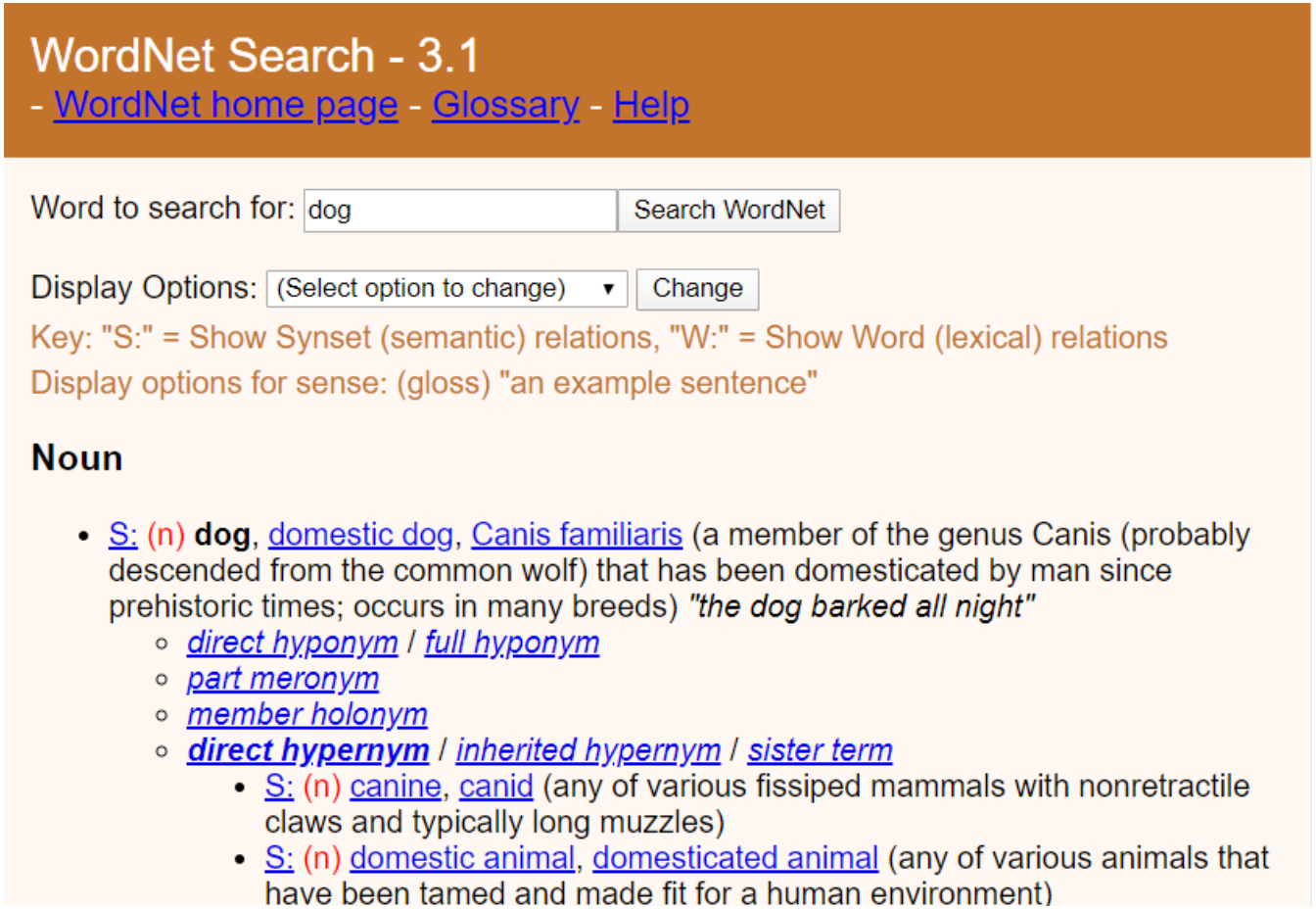
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **plant**, works, industrial plant (buildings for carrying on industrial labor) *"they built a large plant to manufacture automobiles"*
- S: (n) **plant**, flora, plant life ((botany) a living organism lacking the power of locomotion)
- S: (n) **plant** (an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience)
- S: (n) **plant** (something planted secretly for discovery by another) *"the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"*

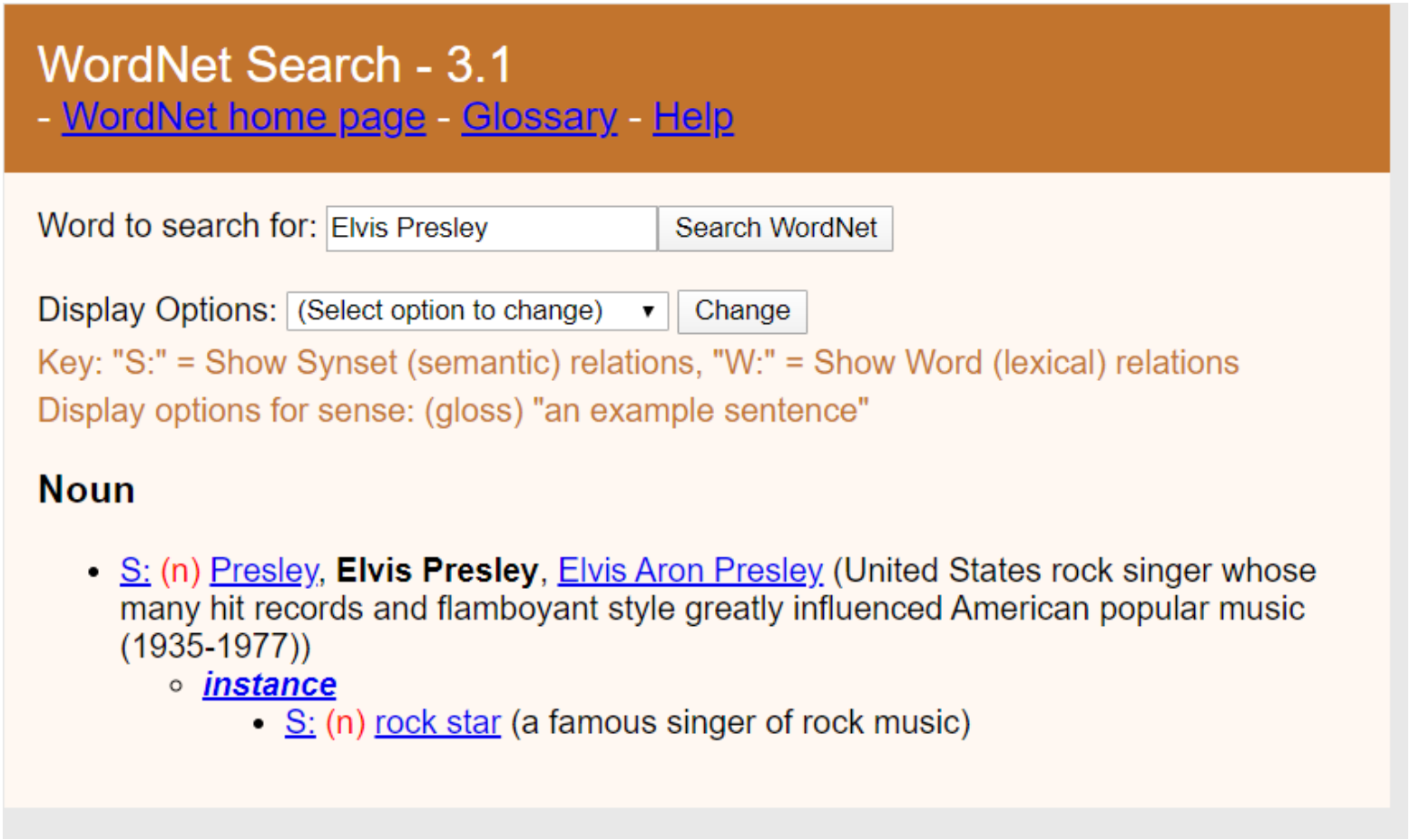
WordNet, hipérnimos



Relaciones en Wordnet

- Hipérnimo / hipónimo (animal, perro)
- Holónimo / merónimo (auto, motor)
- Clase / instancia (Cantante de Rock, Elvis)

WordNet, instancia



WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: Elvis Presley Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) Presley, **Elvis Presley**, Elvis Aron Presley (United States rock singer whose many hit records and flamboyant style greatly influenced American popular music (1935-1977))
  - ***instance***
    - S: (n) rock star (a famous singer of rock music)

Extracción de relaciones: métodos

1) Reglas manuales, patrones

2) Aprendizaje supervisado

3) Semi-supervisado, conjunto semilla

4) Semi-supervisado, supervisión distante

5) Completamente no supervisado, extracción abierta.

Patrones para extracción de relaciones

Método más antiguo (Marti Hearst, 1992)

NP {, NP}* {,} (and|or) other NP_H	temples, treasuries, and other important civic buildings
NP_H such as {NP,}* {(or|and)} NP	red algae such as Gelidium
such NP_H as {NP,}* {(or|and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP_H {,} including {NP,}* {(or|and)} NP	common-law countries, including Canada and England
NP_H {,} especially {NP}* {(or|and)} NP	European countries, especially France, England, and Spain

Figure 18.12 Hand-built lexico-syntactic patterns for finding hypernyms, using {} to mark optionality (Hearst 1992a, Hearst 1998).

La relación es la hipernimia, el hipérnimo está escrito en azul.

Patrones para extracción de relaciones

- Los patrones manuales suelen tener alta precisión pero bajo recall.
- Las distintas reglas pueden interactuar de modo no previsto.
- El agregado de una regla puede interferir con las anteriores.
- Se suelen usar en una 1era etapa, p.ej. para generar un conjunto de ejemplos para supervisión.

Métodos supervisados

En general se procede por etapas, primero las entidades y luego las relaciones.

Tipos de métodos:

Aprendizaje supervisado

La idea es entrenar un clasificador por cada relación, que para cada par de entidades en una oración nos diga si están o no en esa relación (la relación del clasificador, otra opción en un clasificador multiclase)

Atributos:

- Núcleo y tipo de las entidades
- Distancia entre las entidades
- Secuencia de palabras entre las entidades
- **Camino de dependencias entre las entidades**
- Otros

Métodos semi-supervisados

Bootstrapping

Se empieza por un conjunto semilla para una relación R de pares de entidades (o sea, un conjunto de pares)

1. Se encuentran oraciones (web u otro corpus) donde están ambas entidades.

2. Se extrae y generaliza el contexto alrededor de las entidades para extraer nuevos patrones

Métodos semi-supervisados

Bootstrapping

function BOOTSTRAP(*Relation R*) **returns** *new relation tuples*

tuples ← Gather a set of seed tuples that have relation *R*
iterate
sentences ← find sentences that contain entities in *tuples*
patterns ← generalize the context between and around entities in *sentences*
newpairs ← use *patterns* to grep for more tuples
newpairs ← *newpairs* with high confidence
tuples ← *tuples* + *newpairs*
return tuples

Figure 18.15 Bootstrapping from seed entity pairs to learn relations.

Bootstrapping

(AUTOR,LIBRO) — (CORTÁZAR, RAYUELA)

- Hoy os traemos un artículo dedicado a la obra fundamental de **Julio Cortázar, "Rayuela"**.
- En la novela **Rayuela** (1963), **Julio Cortázar** rompe con la concepción tradicional de la narrativa …
- Una de las obras más destacadas de **Cortázar** es **Rayuela**, un clásico de la literatura universal.
- En estos momentos, el juego formal que proponía **Cortáza**r en **Rayuela** puede parecer pretencioso y un tanto innecesario.
- **Rayuela** es una novela del escritor argentino **Julio Cortázar**.

Bootstrapping

(AUTOR,LIBRO) — (CORTÁZAR, RAYUELA)

- Hoy os traemos un artículo dedicado a la obra fundamental de **Julio Cortázar, "Rayuela"**. / obra ? de [autor], [obra] /
- En la novela **Rayuela** (1963), **Julio Cortázar** rompe con la concepción tradicional de la narrativa … /en ? ? [obra] ?, [autor] /
- Una de las obras más destacadas de **Cortázar** es **Rayuela**, un clásico de la literatura universal. /obra ? ? ? de [autor] es [obra] /
- En estos momentos, el juego formal que proponía **Cortáza**r en **Rayuela** puede parecer pretencioso y un tanto innecesario.
- **Rayuela** es una novela del escritor argentino **Julio Cortázar**.

Bootstrapping

Evaluación de un pattern generado

Dado un pattern p, un conjunto actual de pares T1 y de patterns P1 y un conjunto Tn de pares nuevos que satisfacen el pattern

- Productividad (pr), cuántos pares nuevos correctos generó
- Error (er), cuántos pares nuevos incorrectos generó
- Novedad, cuántos de los pares de Tn no se generan por P1 (nov)
- valor(p) ≈ f (nov↑ , er ↓)

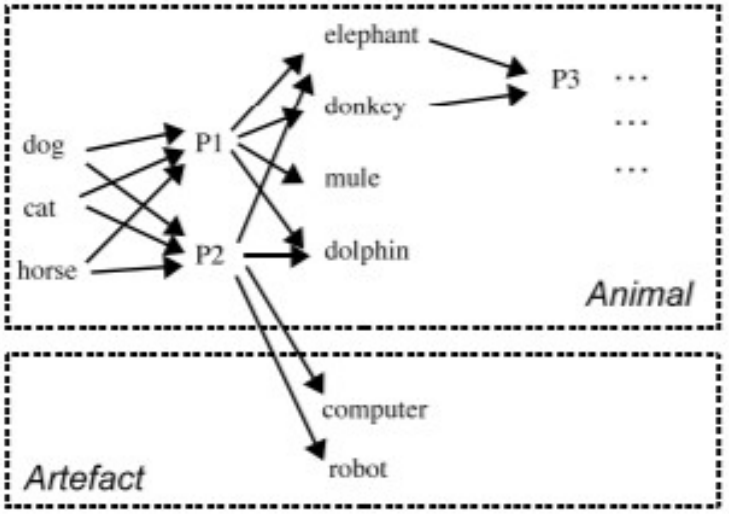
Bootstrapping

Las reglas generadas por el contexto de un par en la relación pueden ocasionar una “mutación” de significado.

Se produce lo que se llama *semantic drift* (alteración del significado).

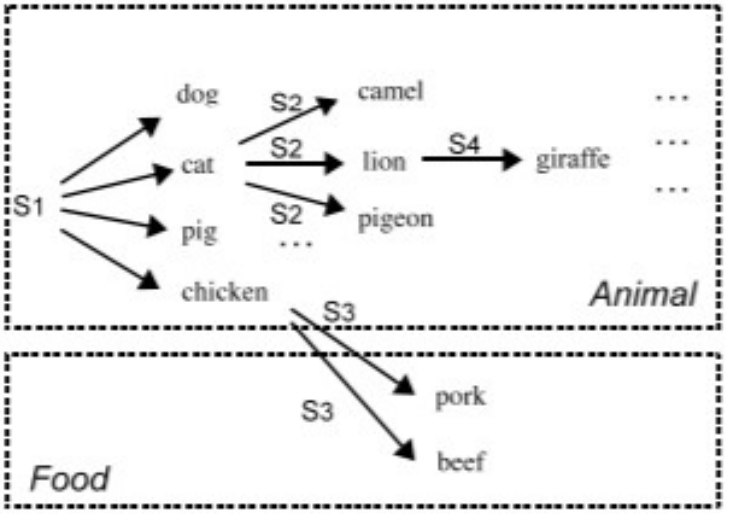
Vemos un ejemplo con 2 tipos de inducción de patterns

Semantic drift - IE



P1: "… X is a kind of mammal …"
P2: "Sometime, X is as clever as human beings"

(a) "syntax-based" bootstrapping mechanism



S1="Animals **such as** dog, cat, pig and chicken, grow fast."
S2="Yoga Postures are named after animals **such as** camel, pigeon, lion and cat."
S3="Common food from animals **such as** pork, beef and chicken."
S4="Animals from African countries **such as** Giraffe and Lion."

(b) "semantic-based" bootstrapping mechanism

Figure 1: Snapshots of Iterative Extraction for "Animal" with Two Different Bootstrapping Mechanisms

Overcoming Semantic Drift in Information Extraction
Zhixu Li , Hongsong Li, Haixun Wang, Yi Yang, Xiangliang Zhang, Xiaofang Zhou ♭ ‡6

Semantic drift – “deriva semántica”

- En el caso de pattern sintáctico, se puede adoptar erróneamente el pattern “ as clever as human being”, que introdujo exitosamente a los delfines pero puede introducir a los robots y a las computadoras y a patterns derivados de ellos.
- El ejemplo con patterns semánticos es similar, por aplicación errónea del pattern se empieza a considerar que *beef* y *pork* son animales

Overcoming Semantic Drift in Information Extraction
Zhixu Li , Hongsong Li, Haixun Wang, Yi Yang, Xiangliang Zhang, Xiaofang Zhou ♭ ‡6

Supervisión distante

- Es también un método semi-supervisado de extracción de relaciones.
- A diferencia de bootstrapping, se supone que se dispone de un conjunto grande de pares en la relación. (Por ej., pares en DBPedia)
- Se buscan instancias de esos pares en la web y se obtiene un conjunto de oraciones que se pueden usar como ejemplos anotados.
- Con un conjunto grande de ejemplos se puede optar por proponer atributos y hacer aprendizaje supervisado, o directamente plantear una red neuronal

Supervisión distante

function DISTANT SUPERVISION(*Database D, Text T*) **returns** *relation classifier C*

foreach relation *R*
  foreach tuple (*e1*,*e2*) of entities with relation *R* in *D*
    sentences ← Sentences in *T* that contain *e1* and *e2*
    f ← Frequent features in *sentences*
    observations ← observations + new training tuple (*e1*, *e2*, *f*, *R*)
  C ← Train supervised classifier on *observations*
  return *C*

Figure 18.16 The distant supervision algorithm for relation extraction. A neural classifier might not need to use the feature set *f*.

Supervisión distante

Ejemplo

Snow et al, 2005, usaron WordNet para extraer hipérnimos

Indujeron los 4 patrones originales de Hearst(1992), además de otros 70.000 :

NPH like NP — Many hormones like leptin...

NPH called NP — using a markup language called XHTML

NP is a NPH — Ruby is a programming language...

NP, a NPH — IBM, a company with a long…

Inconveniente : Se precisa un conjunto grande de pares

Extracción abierta (*OPEN IE*)

No hay datos de entrenamiento.

No hay patrones iniciales.

Solo hay TEXTO !!! Mucho texto

Queremos extraer relaciones, tuplas de relaciones.

Reverb

Usa herramientas de pipeline PLN y heurísticas varias para “adivinar” la expresión de una relación y de sus argumentos.

La intuición para una relación es que está expresada por un verbo (predicado). En realidad, es una unidad verbal multiplabra

*El caballo **trató de pasar** a su contricante.*

*El gobierno **está promoviendo** la exportación en el agro.*

Reverb

Se extraen relaciones.

Se filtran las que no ocurren con al menos 20 argumentos diferentes

Se puntean con un factor de confianza

Ejemplo:

United **has a hub** in Chicago, which **is the headquarters** of United Continental Holdings.

Reverb

- Ejemplo:

  United **has a hub** in Chicago, which **is the headquarters** of United Continental Holdings.

- r1: <United, has a hub in, Chicago>
- r2: <Chicago, is the headquarters of, United Continental Holdings>