Text Mining
Introducción a PLN

Laura Alonso i Alemany

Facultad de Matemática, Astronomía y Física
UNC, Córdoba (Argentina)
http://www.cs.famaf.unc.edu.ar/~laura

SADIO
12 de Marzo de 2008

qué NO es la minería de texto

minería de texto NO es...

- recuperación de información
- aprendizaje automático de ejemplos ya clasificados

objetivos de la minería de texto

- descubrir información desconocida hasta el momento (*knowledge discovery*)
- encontrar patrones para organizar datos

descubrir información desconocida: un ejemplo

Don R. Swanson. 1991. Analysis of Unintended Connections Between Disjoint Science Literatures. SIGIR.

a partir de títulos y abstracts de artículos de medicina se encuentran relaciones causales entre síntomas, drogas, efectos

Problema: dolores de cabeza causados por **migraña**

1. el **estrés** se asocia con la migraña
2. el estrés provoca la pérdida de **magnesio**
   - 2.1 el magnesio es un **bloqueador del canal de calcio** natural
   - 2.2 los bloqueadores del canal de calcio evitan algunas **migrañas**

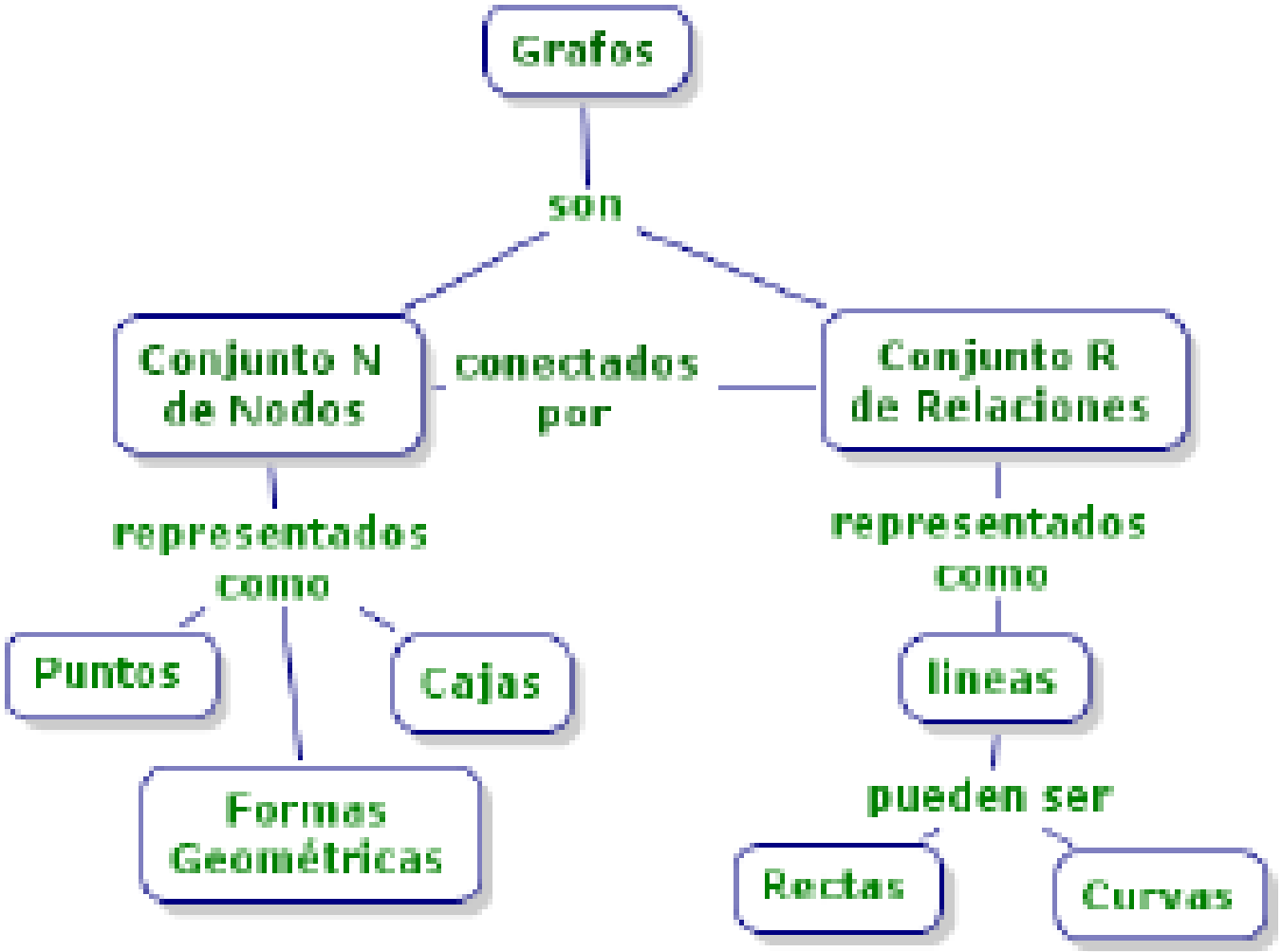
   - 2.1 niveles altos de magnesio inhiben la **depresión cortical extendida**
   - 2.2 la depresión cortical extendida se encuentra en **migrañas**

descubrir patrones para organizar datos: un ejemplo

inducción de mapas conceptuales

descubrir patrones para organizar datos: un ejemplo

inducción de mapas conceptuales



aplicaciones de la minería de texto

minería de texto es...
Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources.

Marti Hearst

- minado de conceptos
- minado de opiniones
- descubrimiento de relaciones
- descubrimiento del comportamiento de las palabras
- descubrimiento de la semántica de las palabras

la frontera es poco clara: qué pasa con los **modelos** que se aprenden de ejemplos clasificados?

limitaciones de la minería de texto

diferencias con KDD tradicional

- casi siempre, ejemplos no clasificados (*unsupervised learning*)
- input **no estructurado**, es necesario
  - identificar entidades
  - identificar relaciones
- rango de resultados **totalmente desconocido**

limitaciones de la minería de texto

- falta de abstracción
  - → alta dimensionalidad
  - → difícil de interpretar
- no está conectado con otros tipos de conocimiento

objetivo del PLN

comprensión automática del lenguaje humano
sub-objetivos:

- desambiguación de sentidos
- análisis sintáctico
- resolución de co-referencia
- interpretación semántica de oraciones

comprensión automática del lenguaje: un ejemplo

sistema de diálogo hombre - máquina

```
H - cuáles son los horarios de los trenes a
Tarragona para mañana?
...
M - a las 7:30, 8, 9, 9:30, 10, 11, 11:30...
```

comprensión automática del lenguaje: un ejemplo

sistema de diálogo hombre - máquina

```
H - cuáles son los horarios de los trenes a
Tarragona para mañana?
...
M - a las 7:30, 8, 9, 9:30, 10, 11, 11:30...
```

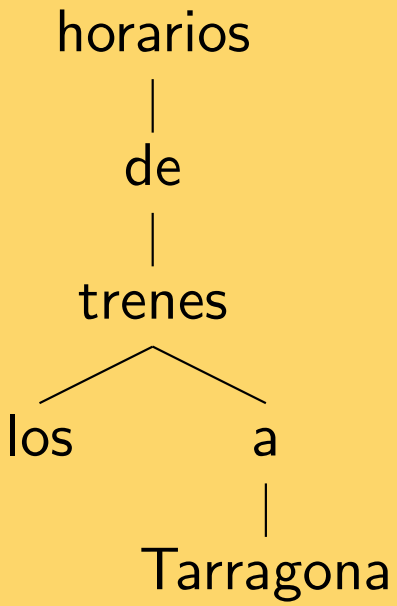
desambiguación de sentidos:
mañana = próximo día
o
mañana = primera parte del día?

comprensión automática del lenguaje: un ejemplo

sistema de diálogo hombre - máquina

```
H - cuáles son los horarios de los trenes a
Tarragona para mañana?
...
M - a las 7:30, 8, 9, 9:30, 10, 11, 11:30...
```

análisis sintáctico:

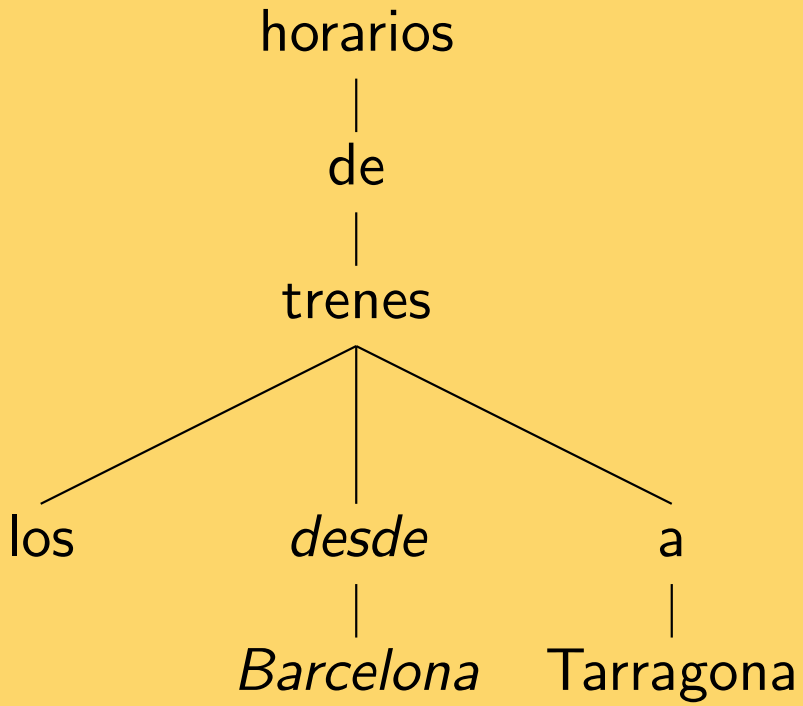


comprensión automática del lenguaje: un ejemplo

sistema de diálogo hombre - máquina

```
H - cuáles son los horarios de los trenes a
Tarragona para mañana?
...
M - a las 7:30, 8, 9, 9:30, 10, 11, 11:30...
```

resolución de co-referencia:



comprensión automática del lenguaje: un ejemplo

sistema de diálogo hombre - máquina

```
H - cuáles son los horarios de los trenes a
Tarragona para mañana?
...
M - a las 7:30, 8, 9, 9:30, 10, 11, 11:30...
```

interpretación semántica de oraciones:

fecha	23/04/2006
medio de transporte	tren
desde	Barcelona-BCN
hasta	Tarragona-TGN
horarios	?

limitaciones del PLN

limitaciones del PLN

cuello de botella: recursos de conocimiento lingüístico

- poca cobertura
- recursos de conocimiento estáticos
- poca adaptación a entornos específicos

cómo ayuda el PLN a la minería de texto

análisis lingüístico

mejor representación del texto

- generalización de fenómenos lingüísticos
  - → se sufre menos la escasez de datos
  - → representación teóricamente más adecuada
- mayor interpretabilidad
- conexión con bases de conocimiento extra-textuales

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
5. análisis semántico

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones

elgatocomepescado

3. análisis morfológico
4. análisis sintáctico
5. análisis semántico

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones

el gato come pescado

3. análisis morfológico
4. análisis sintáctico
5. análisis semántico

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
   - 3.1 detección de palabras especiales

     Woody Allen llegó a Donosti el miércoles a las dos.
   - 3.2 asignación de etiquetas
   - 3.3 desambiguación de etiquetas
4. análisis sintáctico
5. análisis semántico

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
   - 3.1 detección de palabras especiales

Woody Allen llegó a Donosti el miércoles a las dos.

   - 3.2 asignación de etiquetas
   - 3.3 desambiguación de etiquetas
4. análisis sintáctico
5. análisis semántico

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
   - 3.1 detección de palabras especiales
   - 3.2 asignación de etiquetas

el	DA0MS0	el
gato	NCMS000	gato
come	VMIP3S0,VMPP2S0	comer
pescado	NCMS000,VMP00SM	pescado

   - 3.3 desambiguación de etiquetas
4. análisis sintáctico
5. análisis semántico

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
   - 3.1 detección de palabras especiales
   - 3.2 asignación de etiquetas
   - 3.3 desambiguación de etiquetas

el	DA0MS0	el
gato	NCMS000	gato
come	VMIP3S0	comer
pescado	NCMS000	pescado

4. análisis sintáctico
5. análisis semántico

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
   - 4.1 constituyentes básicos o *chunks*

     el gato come pescado
   - 4.2 estructura de oración
   - 4.3 funciones gramaticales, roles temáticos
5. análisis semántico

arquitectura básica de los sistemas de PLN

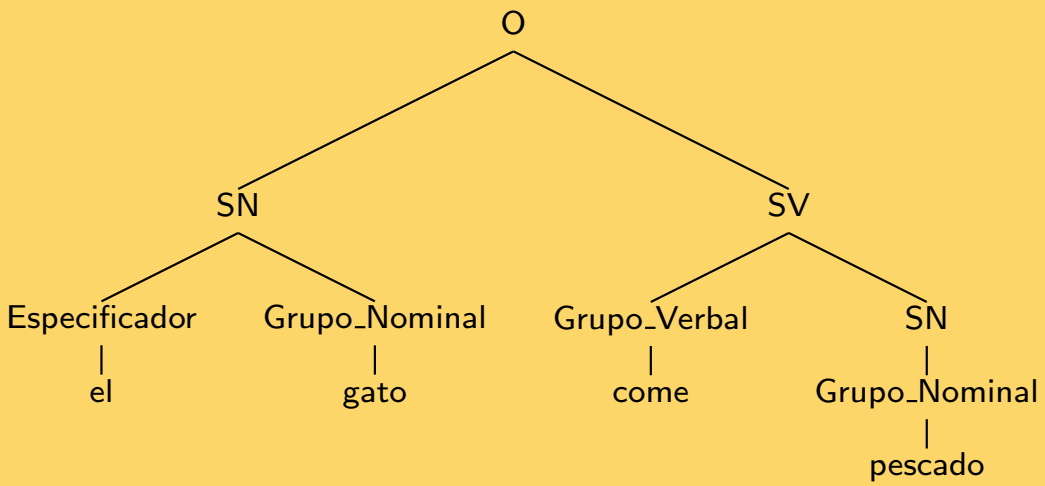
1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
   - 4.1 constituyentes básicos o *chunks*

Grupo_Nominal(el gato) Grupo_Verbal(come) Grupo_Nominal(pescado)

   - 4.2 estructura de oración
   - 4.3 funciones gramaticales, roles temáticos
5. análisis semántico

arquitectura básica de los sistemas de PLN

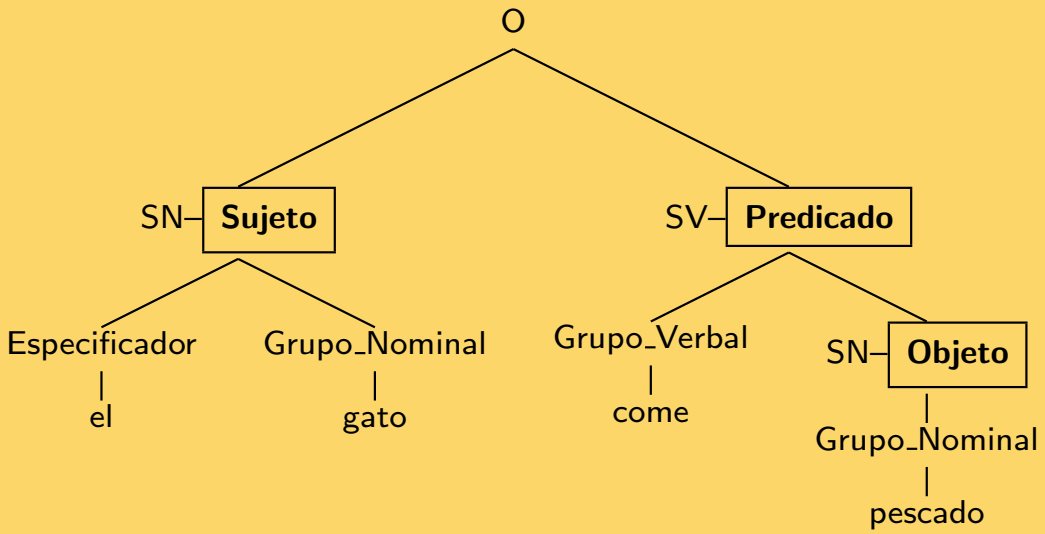
1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
   - 4.1 constituyentes básicos o *chunks*
   - 4.2 estructura de oración



   - 4.3 funciones gramaticales, roles temáticos
5. análisis semántico

arquitectura básica de los sistemas de PLN

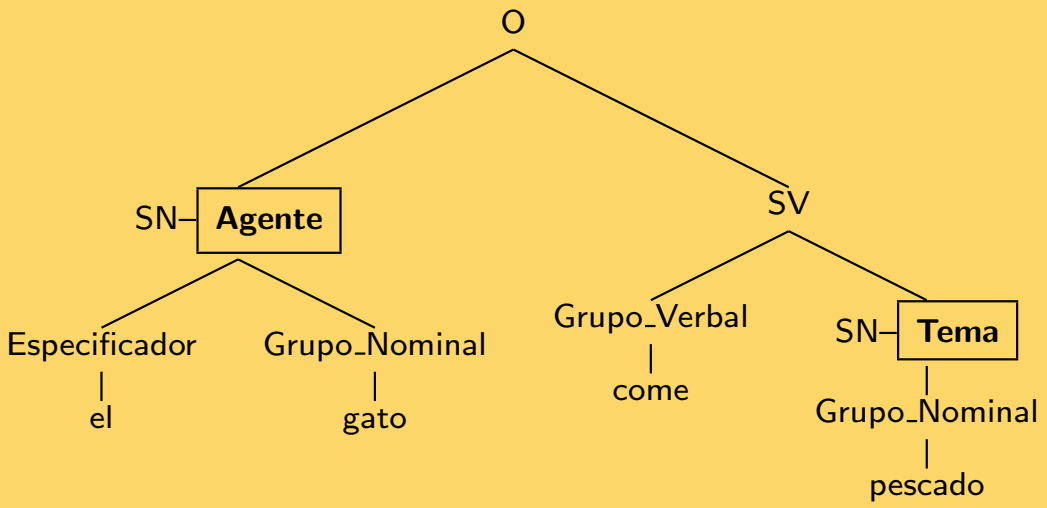
1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
   - 4.1 constituyentes básicos o *chunks*
   - 4.2 estructura de oración
   - 4.3 funciones gramaticales, roles temáticos



5. análisis semántico

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
   - 4.1 constituyentes básicos o *chunks*
   - 4.2 estructura de oración
   - 4.3 funciones gramaticales, roles temáticos



5. análisis semántico

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
   - 4.1 constituyentes básicos o *chunks*
   - 4.2 estructura de oración
   - 4.3 funciones gramaticales, roles temáticos

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
5. análisis semántico
   - 5.1 léxico

el gato	entidad → ser vivo → animal → ... → felino doméstico determinado
come	acción → voluntaria → ...
pescado	entidad → inanimado → natural → comestible entidad → ser vivo → animal → vertebrado → pez no determinado → masa

   - 5.2 proposicional

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
5. análisis semántico
   - 5.1 léxico

Woody Allen	persona → artista → actor → cine persona → artista → director → cine
llegó	acción → desplazamiento → ...
a Donosti	lugar → ciudad
el miércoles a las dos	14:00GMT02/02/2005

   - 5.2 proposicional

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
5. análisis semántico
   - 5.1 léxico
   - 5.2 proposicional

$$\exists \text{gato}(X) \wedge \exists \text{pescado}(Y) \wedge \text{come}(X,Y)$$

arquitectura básica de los sistemas de PLN

1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
5. análisis semántico
   - 5.1 léxico
   - 5.2 proposicional

llega(Woody_Allen,Donosti,14:00GMT02/02/2005)

aproximaciones al PLN

- **arquitecturas simbólicas**
  1. un humano desarrolla reglas de análisis y/o diccionarios
  2. el conocimiento codificado en las reglas y diccionarios se aplica mediante un analizador automático
- **arquitecturas probabilísticas**
  1. uno (o más) humanos analizan una muestra representativa de lenguaje natural (*corpus anotado*)
  2. se aplica un proceso de inferencia de conocimiento (reglas y/o diccionarios) a esta muestra
  3. el conocimiento obtenido automáticamente en forma de reglas y diccionarios, a menudo probabilísticos, se aplica mediante un analizador automático
- **arquitecturas que se ayudan de minería de texto**

cómo ayuda la minería de texto al PLN

- superar el cuello de botella de los recursos de conocimiento
- adaptar los analizadores a la realidad del lenguaje

ejemplos de minería de texto

- técnicas levemente supervisadas
  - aumento de lexicones
  - aumento de ontologías
  - aumento de gramáticas
- técnicas no supervisadas
  - descubrimiento de traducciones
  - descubrimiento de clases de palabras
  - enriquecimiento de lexicones

aumento de lexicones

- reglas morfológicas para crear diccionarios para tagging
- co-ocurrencia con otras palabras para obtener bolsas de palabras temáticas

aumento de lexicones

- **reglas morfológicas para crear diccionarios para tagging**

  Goldsmith, John. 2001. Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics. 27 (2)

  1. se caracterizan los diferentes paradigmas de palabras por sus rasgos morfológicos (`escucho`, `escuchas`, `escuchamos`)
  2. en corpus, se encuentran todas las ocurrencias de una palabra
     - 2.1 se identifica la raíz hipotetizando diferentes opciones dadas las formas encontradas, mediante un árbol de sufijos (trie)
     - 2.2 se asigna cada raíz a uno o más paradigmas, con diferente fiabilidad, usando formas inambiguas del paradigma
     - 2.3 si es necesario, se obtienen más ejemplos (pesados) de la web

  desventajas:
  - hay que construir los paradigmas a mano
  - útil solamente para lenguas con rica morfología flexiva
- co-ocurrencia con otras palabras para obtener bolsas de palabras temáticas

aumento de lexicones

- reglas morfológicas para crear diccionarios para tagging
- **co-ocurrencia con otras palabras para obtener bolsas de palabras temáticas**
  1. partimos de un pequeño diccionario temático (p.ej., Roget's Thesaurus), donde cada tema (*topic*) está asociado a un conjunto de palabras (p.ej., *tenis* → *raqueta, red, set*).
  2. usamos palabras inambiguas (sólo están en una categoría) como indicadoras de tema
  3. buscamos palabras que co-ocurren significativamente más con las palabras inambiguas de un tema (o con documentos asignados a un tema) (p.ej., *Nalbandián*)
  4. incorporamos estas palabras como indicadoras de tema (posiblemente, pesando su indicatividad con su ambigüedad)

útil para categorización de textos, desambiguación de sentidos, inducción de mapas conceptuales

desventajas: es necesario tener un inventario de temas pre-definido

aumento de lexicones

- reglas morfológicas para crear diccionarios para tagging
- **co-ocurrencia con otras palabras para obtener bolsas de palabras temáticas**
  1. partimos de un pequeño diccionario temático (p.ej., Roget's Thesaurus), donde cada tema (*topic*) está asociado a un conjunto de palabras (p.ej., *tenis* → *raqueta, red, set*).
  2. usamos palabras inambiguas (sólo están en una categoría) como indicadoras de tema
     alternativa: en un documento las diferentes palabras votan por sus diferentes temas (repartiendo su capacidad de voto equitativamente entre sus diferentes temas) para determinar el tema del documento
  3. buscamos palabras que co-ocurren significativamente más con las palabras inambiguas de un tema (o con documentos asignados a un tema) (p.ej., *Nalbandián*)
  4. incorporamos estas palabras como indicadoras de tema (posiblemente, pesando su indicatividad con su ambigüedad)

útil para categorización de textos, desambiguación de

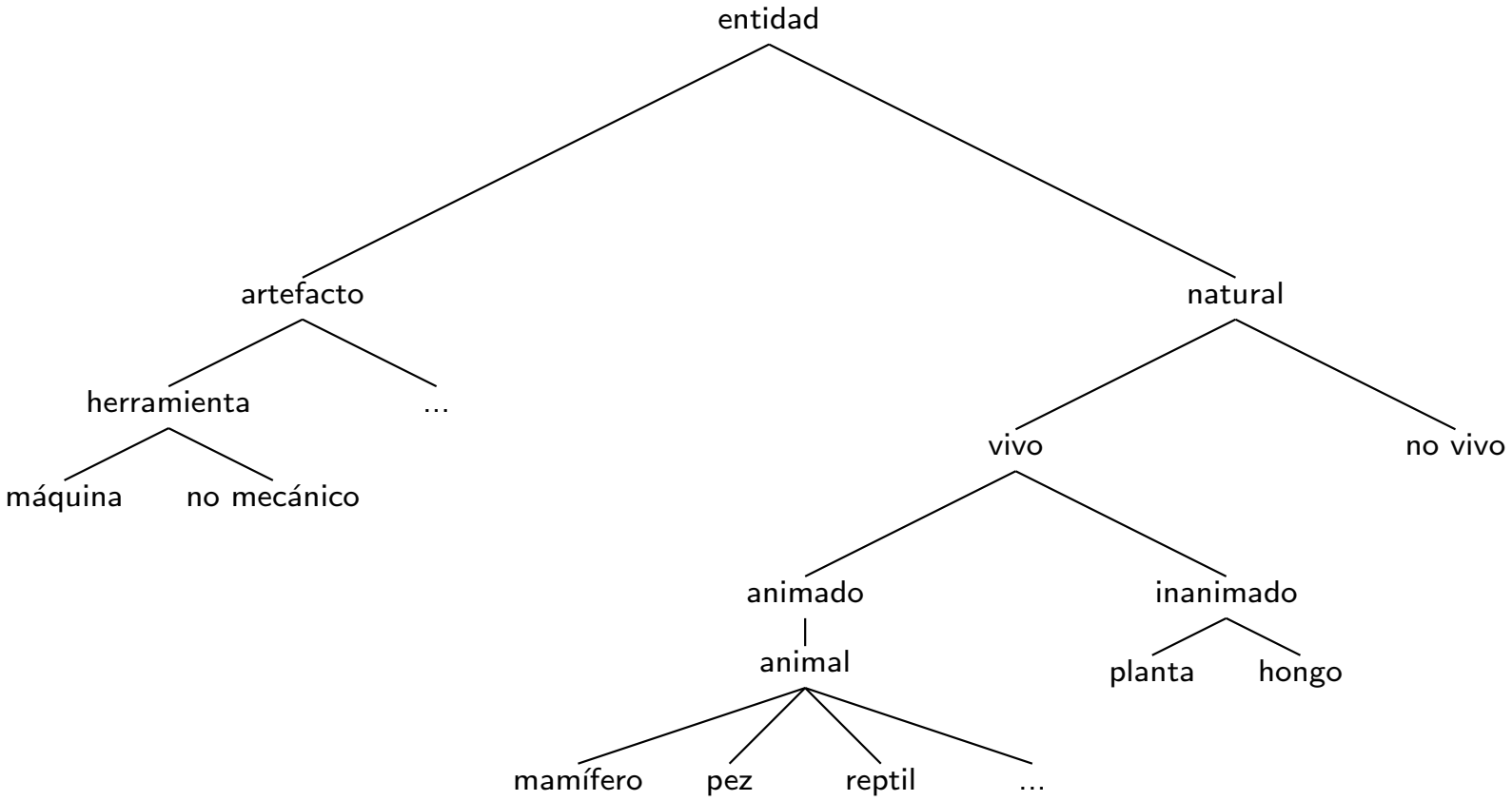
aumento de lexicones

- reglas morfológicas para crear diccionarios para tagging
- **co-ocurrencia con otras palabras para obtener bolsas de palabras temáticas**
  1. partimos de un pequeño diccionario temático (p.ej., Roget's Thesaurus), donde cada tema (*topic*) está asociado a un conjunto de palabras (p.ej., *tenis* → *raqueta, red, set*).
  2. usamos palabras inambiguas (sólo están en una categoría) como indicadoras de tema
  3. buscamos palabras que co-ocurren significativamente más con las palabras inambiguas de un tema (o con documentos asignados a un tema) (p.ej., *Nalbandián*)
  4. incorporamos estas palabras como indicadoras de tema (posiblemente, pesando su indicatividad con su ambigüedad)

útil para categorización de textos, desambiguación de sentidos, inducción de mapas conceptuales

desventajas: es necesario tener un inventario de temas pre-definido

aumento de ontologías



la ontología estrella para PLN es WordNet (y sus extensiones: EuroWordNet, BalkaNet, GermaNet...)

aumento de ontologías

- patrones sintácticos para descubrir hipónimos, hiperónimos
- co-ocurrencia con otras palabras para determinar adjunción a una ontología

aumento de ontologías

- **patrones sintácticos para descubrir hipónimos, hiperónimos**
  `N es un N, N es un tipo de N, N pertenece a los N`
  (*el tigre es un felino, Rosina es la presidenta de la Liga de Madres*)
  útil si tenemos un corpus donde estos patrones se dan significativamente: diccionario, textos didácticos
  desventajas: alcance limitado, conocimiento típicamente enciclopédico (vs. sentido común), multiplicidad de patrones posiblemente ambiguos
- **co-ocurrencia con otras palabras para determinar adjunción a una ontología**

aumento de ontologías

- patrones sintácticos para descubrir hipónimos, hiperónimos
- co-ocurrencia con otras palabras para determinar adjunción a una ontología
  - usando técnicas basadas en desambiguación de sentidos
  - usando técnicas basadas en tematización

aumento de ontologías

- patrones sintácticos para descubrir hipónimos, hiperónimos
- **co-ocurrencia con otras palabras para determinar adjunción a una ontología**
  - **usando técnicas basadas en desambiguación de sentidos**
    dada una palabra ambigua *p* de la ontología,
    1. para todos los sentidos de *p*, buscamos otras palabras que estén asignadas únicamente a uno de sus sentidos (o hipo-sentidos, o hiper-sentidos)
    2. recuperamos de la web la co-ocurrencia de *p* con cada una de las palabras inambiguas
    3. pesamos los sentidos de *p* por su co-ocurrencia con estas palabras inambiguas

    desventajas: la palabra tiene que estar previamente descrita en la ontología, escasez de palabras inambiguas
  - **usando técnicas basadas en tematización**

aumento de ontologías

- patrones sintácticos para descubrir hipónimos, hiperónimos
- **co-ocurrencia con otras palabras para determinar adjunción a una ontología**
  - **usando técnicas basadas en desambiguación de sentidos**
  - **usando técnicas basadas en tematización**
    1. caracterizamos una rama de la ontología por las palabras que se encuentran en ella, pesando la contribución de su palabra por su ocurrencia en otras ramas
    2. dadas las palabras de un documento que se encuentran en la ontología, determinamos la rama mayoritaria de un documento
    3. encontramos las palabras que no están en la ontología y que ocurren significativamente más con una rama que con el resto, y las incorporamos a la ontología

  desventajas: las diferentes ramas de una ontología pueden co-existir en un mismo documento, cobertura desigual de las ontologías

aumento de gramáticas

- **bootstrapping**
  1. pequeño recurso inicial (gramática o corpus)
  2. se analiza un corpus mayor
  3. se toman en cuenta los ejemplos con mayor fiabilidad
  4. se aprende una gramática mayor
- **co-training**
  1. pequeño recurso inicial (gramática o corpus)
  2. se aprenden dos gramáticas diferentes
  3. se analiza un corpus mayor
  4. se toman en cuenta los ejemplos analizados igual por las dos gramáticas
  5. se aprenden gramáticas mayores

descubrimiento de traducciones

Traducción Automática Estadística

1. corpus paralelos (corpus con el mismo significado en dos lenguas distintas, p.ej., corpus del Parlamento Europeo)
2. alineación: identificar correspondencias entre segmentos cada vez más chicos de los dos textos
3. correspondencias entre palabras o secuencias de cada lengua:
   - palabra - palabra
   - palabra - $\emptyset$
   - $\emptyset$ - palabra
   - palabra palabra - palabra
   - palabra palabra - palabra palabra
   - ...
4. asignamos una fiabilidad a cada correspondencia, según la cantidad de veces que la hemos visto en el corpus
5. interpolación de modelos de diferente longitud y fiabilidad

descubrimiento de clases de palabras

- caracterizar cada palabra p como un vector n-dimensional,
- cada dimensión se corresponde con una de las palabras que ocurren en el texto,
- la dimensión n del vector que caracteriza a la palabra p contiene la probabilidad de co-ocurrencia de la palabra p con la palabra de la dimensión n.

descubrimiento de clases de palabras

El gato come pescado. El pescado vive en el mar. El mar está lleno de agua. A los gatos no les gusta el agua. A los gatos les gusta el pescado.

palabras a caracterizar:

```
A El agua.  come de el en está gato gatos gusta les
lleno los mar mar.  no pescado pescado.  vive
```

descubrimiento de clases de palabras

El gato come pescado. El pescado vive en el mar. El mar está lleno de agua. A los gatos no les gusta el agua. A los gatos les gusta el pescado.

palabras a caracterizar: **-10%**

A El agua come de el en está gato gatos gusta les lleno los mar no pescado vive

descubrimiento de clases de palabras

El gato come pescado. El pescado vive en el mar. El mar está lleno de agua. A los gatos no les gusta el agua. A los gatos les gusta el pescado.

palabras a caracterizar: **-15%**

a agua come de el en está gato gatos gusta les lleno los mar no pescado vive

descubrimiento de clases de palabras

El gato come pescado. El pescado vive en el mar. El mar está lleno de agua. A los gatos no les gusta el agua. A los gatos les gusta el pescado.

palabras a caracterizar: **-20%**

`a agua come de el en está gato gusta les lleno los mar no pescado vive`

descubrimiento de clases de palabras

El gato come pescado. El pescado vive en el mar. El mar está lleno de agua. A los gatos no les gusta el agua. A los gatos les gusta el pescado.

palabras a caracterizar: **-60%**

`agua come gato gusta lleno mar pescado vive`

descubrimiento de clases de palabras

El gato come pescado. El pescado vive en el mar. El mar está lleno de agua. A los gatos no les gusta el agua. A los gatos les gusta el pescado.

palabras a caracterizar:

`agua come gato gusta lleno mar pescado vive`

	agua	come	gato	gusta	lleno	mar	pescado	vive
agua			1	1	1	1		
come			1				1	
gato	1	1		2			2	
gusta	1		2				1	
lleno	1					1		
mar	1				1		1	1
pescado		1	2	1		1		1
vive						1	1	

descubrimiento de clases de palabras

	agua	come	gato	gusta	lleno	mar	pescado	vive
agua			1	1	1	1		
come			1				1	
gato	1	1		2			2	
gusta	1		2				1	
lleno	1					1		
mar	1				1		1	1
pescado		1	2	1		1		1
vive						1	1	

usar técnicas de *clustering* para encontrar grupos de significado:

determinar qué vectores están más cercanos,
según un criterio de distancia matemático

descubrimiento de clases de palabras

	agua	come	gato	gusta	lleno	mar	pescado	vive
agua			1	1	1	1		
come			1				1	
gato	1	1		2			2	
gusta	1		2				1	
lleno	1					1		
mar	1				1		1	1
pescado		1	2	1		1		1
vive						1	1	

usar técnicas de *clustering* para encontrar grupos de significado:

determinar qué **palabras** están más cercanas,
según un criterio de distancia matemático

descubrimiento de clases de palabras

	agua	come	gato	gusta	lleno	mar	pescado	vive
agua			1	1	1	1		
come			1				1	
gato	1	1		2			2	
gusta	1		2				1	
lleno	1					1		
mar	1				1		1	1
pescado		1	2	1		1		1
vive						1	1	

usar técnicas de *clustering* para encontrar grupos de significado:

determinar qué **palabras se parecen más entre sí**,
según un criterio de distancia matemático

cuestiones a tener en cuenta en clustering

- selección de características (las dimensiones de los vectores)
  - ni muchas, ni pocas
  - adecuadas al objetivo que pretendemos conseguir.
- trabajar suficientes instancias de cada objeto para obtener representatividad estadística
- selección del criterio de distancia entre vectores
  - **distancia euclídea**: el más simple
  - **coseno**: el más usado, con los mejores resultados
  - **divergencia de Kullback-Leibler**, no es una distancia sino una diferencia entre distribuciones de probabilidad
- también se pueden usar técnicas más sofisticadas, como la descomposición en valores singulares, desarrollada en la técnica de Latent Semantic Analysis (`T. K. Landauer, P. W. Foltz, and D. Laham. 1998. Introduction to Latent Semantic Analysis. Discourse Processes, 25.`)

enriquecimiento de lexicones

problema inicial: la información de los lexicones computacionales es insuficiente para aplicaciones de conocimiento profundo
objetivo: asociar a los ítems de los lexicones información sobre su comportamiento respecto a las palabras circundantes

1. obtener ocurrencias en corpus de las palabras que queremos enriquecer (p.ej., verbos), posiblemente analizadas (p.ej., análisis sintáctico, semántico)
2. caracterizar las palabras-objetivo como vectores
3. aplicar técnicas de clustering o bootstrapping para encontrar clases de palabras con el mismo comportamiento sintáctico-semántico
4. aprender un clasificador a partir de estas clases y estos ejemplos
5. aplicarlo a palabras semejantes pero antes no vistas