Taller de Métodos y Herramientas de Procesamiento del Lenguaje Natural

Laura Alonso i Alemany

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

ELiC
26, 27 y 28 de Julio de 2010

grupo de PLN en FaMAF

`http://www.cs.famaf.unc.edu.ar/~pln/`

- Doctores
  - Gabriel Infante-Lopez – análisis sintáctico, theorem provers
  - Laura Alonso Alemany – métodos empíricos, subcategorización
  - Paula Estrella – traducción automática
  - Luciana Benotti – generación de lenguaje natural, interacción persona-computador
  - Carlos Areces – generación de lenguaje natural, lógicas híbridas
- Estudiantes de Doctorado
  - Martín Domínguez – análisis sintáctico
  - Franco Luque – análisis sintáctico
  - Romina Altamirano – reconocimiento de implicación textual
  - Raúl Fervari – lógicas y ontologías
  - Santiago Correa – minería en texto semi-supervisado

objetivos de la ELiC

- presentar la Lingüística Computacional a estudiantes de la Argentina y del resto de Latinoamérica,
- ofrecer una perspectiva general y actualizada del área, mostrando diferentes alternativas para atacar problemas de Lingüística Computacional y su relación con propuestas teóricas y aplicaciones prácticas.
- familiarizar a los asistentes con problemas y soluciones de los temas del área,
- servir de punto de encuentro para investigadores de Lingüística Computacional.
- contribuir a la creación de una comunidad latinoamericana de investigación en el área.

pero qué es la Lingüística Computacional?

La **Lingüística Computacional** (también conocida como Procesamiento del Lenguaje Natural, *PLN*) es un área **interdisciplinaria** de conocimiento

pero qué es la Lingüística Computacional?

La **Lingüística Computacional** (también conocida como Procesamiento del Lenguaje Natural, *PLN*) es un área **interdisciplinaria** de conocimiento que propone **modelos computacionales** del **lenguaje humano**,

pero qué es la Lingüística Computacional?

La **Lingüística Computacional** (también conocida como Procesamiento del Lenguaje Natural, *PLN*) es un área **interdisciplinaria** de conocimiento que propone **modelos computacionales** del **lenguaje humano**, modelos que pueden ser **basados en reglas** o bien **estadísticos**.

pero qué es la Lingüística Computacional?

La **Lingüística Computacional** (también conocida como Procesamiento del Lenguaje Natural, *PLN*) es un área **interdisciplinaria** de conocimiento que propone **modelos computacionales** del **lenguaje humano**, modelos que pueden ser **basados en reglas** o bien **estadísticos**.
La Lingüística Computacional trata muchos aspectos del lenguaje humano: desde el habla espontánea hasta el texto escrito, pasando por manuales técnicos, noticias de diarios o informes médicos.

pero qué es la Lingüística Computacional?

La **Lingüística Computacional** (también conocida como Procesamiento del Lenguaje Natural, *PLN*) es un área **interdisciplinaria** de conocimiento que propone **modelos computacionales** del **lenguaje humano**, modelos que pueden ser **basados en reglas** o bien **estadísticos**.
La Lingüística Computacional trata muchos aspectos del lenguaje humano: desde el habla espontánea hasta el texto escrito, pasando por manuales técnicos, noticias de diarios o informes médicos.
Sus objetivos van desde crear modelos teóricos sobre la organización y evolución del lenguaje, hasta desarrollar aplicaciones como la traducción automática o los sistemas de diálogo telefónicos.

objetivos de este curso

- presentar aplicaciones habituales y futuras de la Lingüística Computacional
- familiarizar con los problemas básicos del PLN
- presentar diferentes aproximaciones para la resolución de problemas
- presentar las herramientas disponibles para resolver algunos de los problemas de PLN
- motivar la necesidad de evaluación de los sistemas de PLN
- presentar diferentes metodologías de evaluación
- capacitar a los estudiantes para implementar sus propias soluciones

contenidos del curso

- día 1
- día 2
- día 3
- día 4
- día 5
- día 6

contenidos del curso

- día 1
  - objetivos del PLN
  - aplicaciones
  - niveles de análisis del lenguaje
  - algunas herramientas
- día 2
- día 3
- día 4
- día 5
- día 6

contenidos del curso

- día 1
- día 2
  - presentación y uso de FreeLing
  - presentación y uso de Weka
- día 3
- día 4
- día 5
- día 6

contenidos del curso

- día 1
- día 2
- día 3
  - recursos lingüísticos semánticos: EuroWordNet, estructura y extensiones, otras ontologías y bases de conocimiento
  - métodos no supervisados y semisupervisados en PLN
  - clustering para descubrir clases de palabras
- día 4
- día 5
- día 6

contenidos del curso

- día 1
- día 2
- día 3
- día 4 Paula Estrella y el fabuloso mundo de la evaluación
- día 5 Paula Estrella y el fabuloso mundo de la evaluación
- día 6

contenidos del curso

- día 1
- día 2
- día 3
- día 4
- día 5
- día 6 evaluación! *take home*

contenidos del primer día

qué es el PLN

arquitectura de sistemas de PLN

herramientas

- preprocesos
- análisis morfológico (*tagging*)
- análisis sintáctico superficial (*chunking*)
- análisis sintáctico (*parsing*)
- análisis semántico

otros recursos

- aplicaciones
- cajas de herramientas
- directorios de herramientas, recursos y documentación

tarea para el hogar

aplicaciones de PLN

- recuperación de información
- acceso a bases de datos en lenguaje natural
- corrección automática (y sugerencia de palabras)
- resumen automático
- traducción automática
- soporte al aprendizaje de lenguas por computadora
- soporte a la descripción de lenguas por computadora
- creación semiautomática de mapas conceptuales
- detección de sentimientos

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...
```

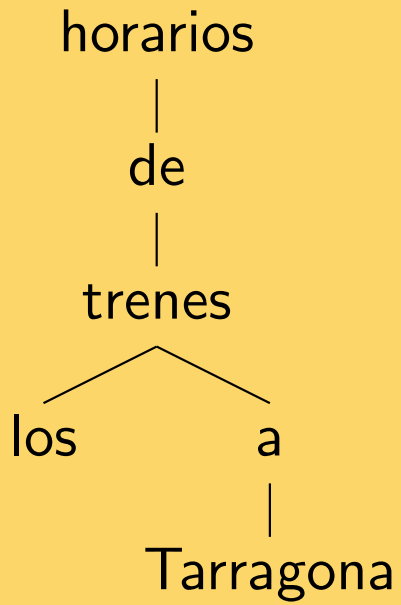
desambiguación de sentidos:
mañana = próximo día
o
mañana = primera parte del día?

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:

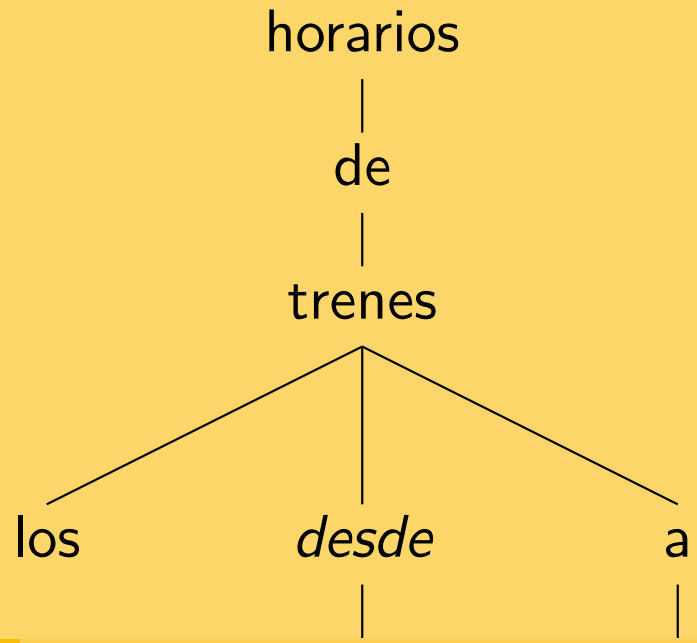


comprensión automática del lenguaje: un ejemplo

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

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

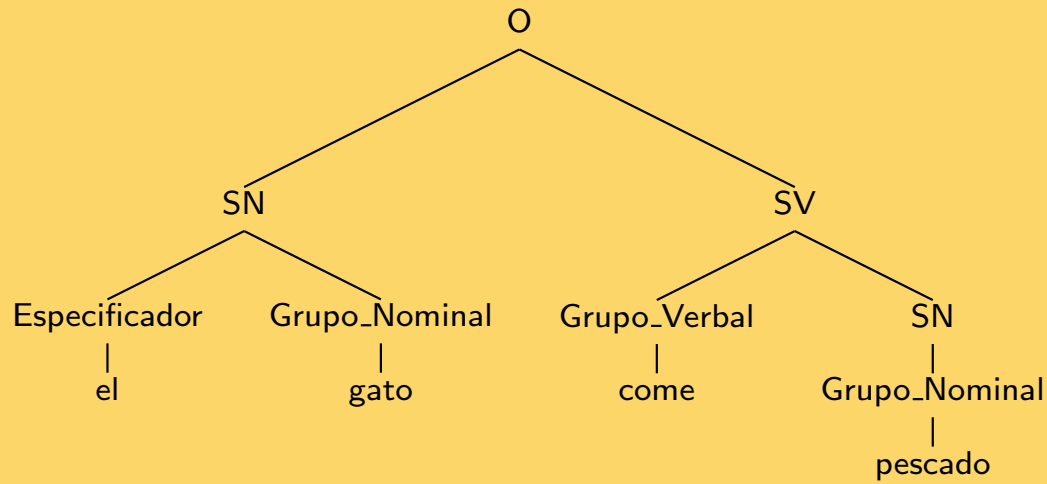
1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
   - 4.1 constituyentes básicos o *chunks*

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

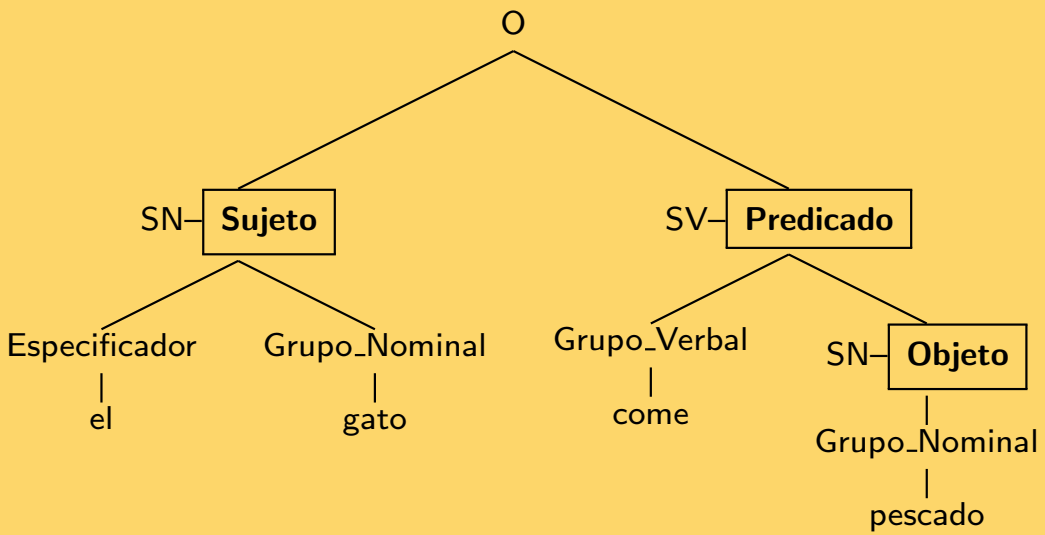
1. reconocimiento de idioma
2. segmentación de palabras, oraciones, secciones
3. análisis morfológico
4. análisis sintáctico
   - 4.1 constituyentes básicos o *chunks*
   - 4.2 estructura de oración



   - 4.3 funciones gramaticales, roles temáticos
5. análisis semántico

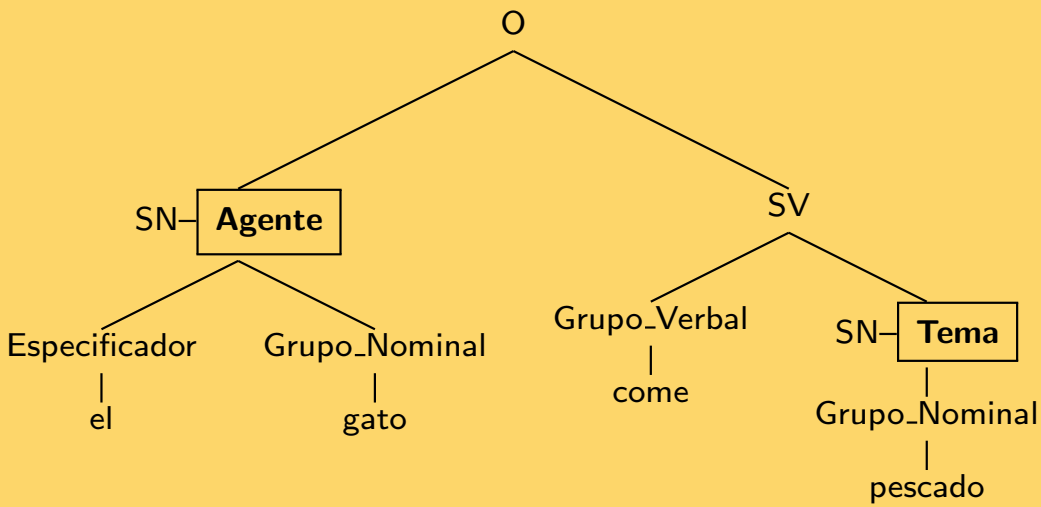
arquitectura básica de los sistemas de PLN

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



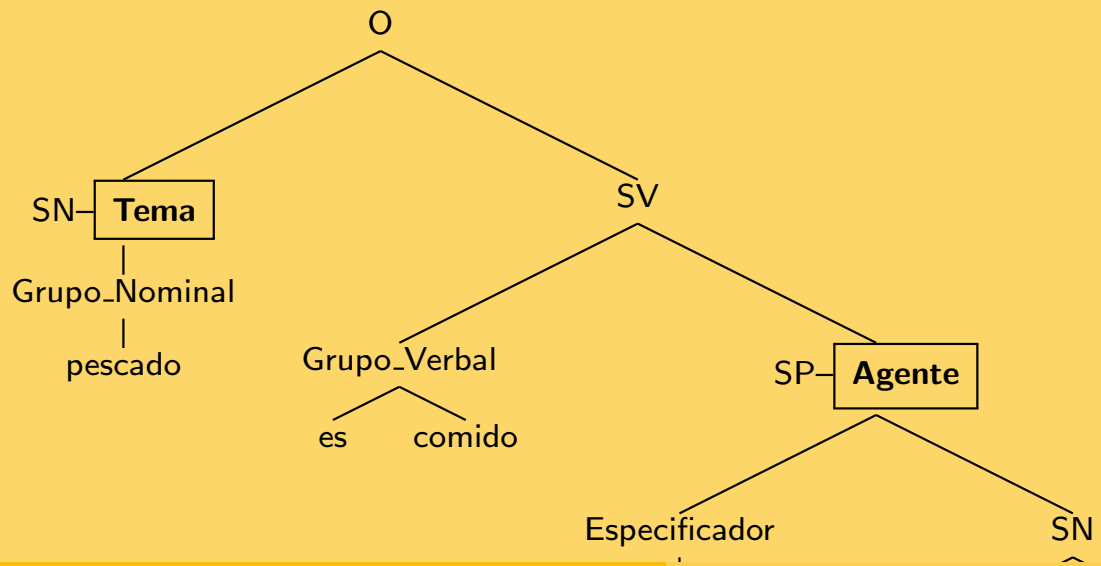
arquitectura básica de los sistemas de PLN

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
   - 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 basadas en conocimiento**
  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 basadas en aprendizaje automático**
  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 menudo probabilísticos) a esta muestra
  3. el conocimiento obtenido automáticamente se aplica mediante un analizador automático
- **arquitecturas basadas en aprendizaje no supervisado**
  1. se aplica un proceso de detección de patrones significativos a un corpus de la lengua suficientemente grande
  2. los patrones (posiblemente validados por un experto humano) se aplican a tareas de PLN

analizadores multi-nivel

- FreeLing
- NLTK
- LingPipe
- OpenNLP

analizadores multi-nivel

- FreeLing código abierto (LGPL), con diccionarios y gramáticas para español, catalán, galaico-portugués, asturianu, inglés, italiano, galés... Tiene un diccionario del español que cubre más del 90% de la lengua, el diccionario de más cobertura de uso totalmente libre. Puede hacer:
  - WN-based semantic information access
  - UKB word sense disambiguation.
  - More expressive rule language for dependency parsing
  - Coreference resolution
  - Machine Learning functionalites moved to external omlet+fries library, for clearer organization
  - Suport for 64-bit processors
  - Extended Java API
- NLTK
- LingPipe
- OpenNLP

analizadores multi-nivel

- FreeLing
- NLTK Open source Python modules, linguistic data and documentation for research and development in natural language processing and text analytics, with distributions for Windows, Mac OSX and Linux.
- LingPipe
- OpenNLP

analizadores multi-nivel

- FreeLing
- NLTK
- LingPipe
- OpenNLP

analizadores multi-nivel

- FreeLing
- NLTK
- LingPipe
- OpenNLP

preprocesos para el análisis

- identificación de lengua
- segmentación de palabras (*tokenization*), oraciones, párrafos, secciones
- lematización (o *stemming*)

preprocesos para el análisis

- identificación de lengua
- segmentación de palabras (*tokenization*), oraciones, párrafos, secciones
  - muy interesante: el trabajo de segmentación de oraciones de Advait Ratnaparkhi, como ejemplo de aplicación de modelos de máxima entropía (y los ejercicios asociados en el libro de Manning y Schütze)
  - el sentence splitter un divisor en oraciones heurístico, implementado en Java, obtiene 99% de fiabilidad en detectar oraciones y párrafos en texto biomédico del inglés.
  - LT-TTT2 componentes para tokenización, sentences splitting, chunking e identificación de entidades nombradas basada en reglas.
- lematización (o *stemming*)

identificación de palabras especiales

son palabras especiales todo lo que no aparece en los diccionarios

- términos
- entidades nombradas

identificación de palabras especiales

son palabras especiales todo lo que no aparece en los diccionarios

- términos → extracción de terminología
- entidades nombradas (NERC *Named Entity Recognition and Classification*) nombres de personas, entidades, lugares, eventos... y también fechas, cifras.

identificación de palabras especiales

son palabras especiales todo lo que no aparece en los diccionarios

- términos
- entidades nombradas

detectores de entidades nombradas

- los analizadores morfosintácticos más completos incluyen utilidades NERC, usando listas y heurísticas simples.
- C&C NER
- Named Entity Tagger from Cognitive Computation Group at the University of Urbana-Champaign
- un programa de perl que cubre muchos de los tipos de expresiones temporales contempladas en las directrices de anotación de TIMEX2
- Acrophile Tratamiento de acrónimos y siglas, diccionarios de acrónimos, identificación y extracción

el análisis morfológico

la mayor parte de herramientas asignan y desambiguan a la vez, y todas incluyen lematización

1. asignación de etiquetas
2. desambiguación de etiquetas

el análisis morfológico

la mayor parte de herramientas asignan y desambiguan a la vez, y todas incluyen lematización

1. asignación de etiquetas
2. desambiguación de etiquetas

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

el análisis morfológico

la mayor parte de herramientas asignan y desambiguan a la vez, y todas incluyen lematización

1. asignación de etiquetas
2. desambiguación de etiquetas

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

diccionarios de palabras

- todos los analizadores morfológicos y sintácticos tienen un diccionario, en los casos de analizadores de código abierto, el diccionario es accesible
- para la lengua castellana, el diccionario de Freeling cubre un 90% de la lengua
- un importantísimo recurso léxico es WordNet y sus extensiones (EuroWordNet, BalkaNet y muchos otros), que veremos en la parte de análisis semántico.
- la lista de lemarios del castellano de Ismael Olea no tiene desperdicio.

correctores ortográficos

- Xuxen es un corrector ortográfico para el vasco
- ispell International Ispell is an interactive spell-checking program for Unix which supports a large number of European languages. An emacs interface is available as well as the standard command-line mode.
- aspell GNU Aspell is a Free and Open Source spell checker designed to eventually replace Ispell.
- el diccionario para el español COES está integrado en ispell y es de esperar que pronto lo esté en aspell.
- myspell es el corrector ortográfico de OpenOffice, basado en aspell.

cómo se obtienen *taggers* probabilísticos

corpus de entrenamiento

el	*gato*	*come*	*pescado*
DA0MS0	*NCMS000*	*VMIP3S0*	*NCMS000*

cómo se obtienen *taggers* probabilísticos

corpus de entrenamiento

método de inferencia

modelos ocultos de Markov (HMM),
modelos de máxima entropía, y otros

cómo se obtienen *taggers* probabilísticos

corpus de entrenamiento | **método de inferencia** | **herramienta de análisis**

el-DA0MS0 gato-NCMS0 **come VMIP3S0 VMPP2S0**

cómo se obtienen *taggers* probabilísticos

corpus de entrenamiento

método de inferencia

herramienta de análisis

el-DA0MS0 gato-NCMS0 **come VMIP3S0 VMPP2S0**
$\xrightarrow{\text{analizador}}$ come-VMIP3S0

corpus anotados

para el español: AnCora-ES un corpus con 188.513 palabras en 6.009 oraciones, anotadas manualmente con funciones sintácticas y synsets nominales de WordNet. También para el catalán (395.379 words in 13.712 sentences)

corpus anotados

para otras lenguas:

- Susanne es un extracto de 130.000 palabras del corpus Brown de inglés americano, analizadas sintácticamente
- Christine es un extracto de 80.000 palabras de lenguaje oral del corpus del inglés British National Corpus, analizadas sintácticamente
- Lucy es un corpus del inglés británico de 165.000 palabras, analizadas sintácticamente
- Cambridge's evaluation corpus for English Subcategorization Acquisition Systems un corpus de 65.000 palabras para evaluar sistemas de adquisición de subcategorización verbal para 30 verbos, con 1000 ocurrencias para cada verbo.

corpus anotados

el repositorio de corpus de referencia es pago (y mucho!):
Linguistic Data Consortium

analizadores morfológicos

- Stanford POS tagger java, código abierto (GPL). Se incluyen dos modelos para el inglés.
- Brill's Transformation-based learning Tagger
- Maximum Entropy part of speech tagger MXPOST
- TnT
- SVMTool
- TreeTagger
- Original Xerox Tagger
- μ-TBL
- QTA
- Lingua-EN-Tagger
- PoSTech Korean morphological analyzer and tagger

analizadores morfológicos

- Stanford POS tagger
- Brill's Transformation-based learning Tagger C, aproximación simbólica.
- Maximum Entropy part of speech tagger MXPOST
- TnT
- SVMTool
- TreeTagger
- Original Xerox Tagger
- μ-TBL
- QTA
- Lingua-EN-Tagger
- PoSTech Korean morphological analyzer and tagger

analizadores morfológicos

- Stanford POS tagger
- Brill's Transformation-based learning Tagger
- Maximum Entropy part of speech tagger MXPOST java (Archivos de clases, no fuente). Incluye un detector de finales de oración.
- TnT
- SVMTool
- TreeTagger
- Original Xerox Tagger
- μ-TBL
- QTA
- Lingua-EN-Tagger
- PoSTech Korean morphological analyzer and tagger

analizadores morfológicos

- Stanford POS tagger
- Brill's Transformation-based learning Tagger
- Maximum Entropy part of speech tagger MXPOST
- TnT para Solaris y Linux. Muy eficiente. Incluye modelos para inglés y alemán. Licencia de uso libre para fines no comerciales.
- SVMTool
- TreeTagger
- Original Xerox Tagger
- μ-TBL
- QTA
- Lingua-EN-Tagger
- PoSTech Korean morphological analyzer and tagger

analizadores morfológicos

- Stanford POS tagger
- Brill's Transformation-based learning Tagger
- Maximum Entropy part of speech tagger MXPOST
- TnT
- SVMTool C y Perl, código abierto (LGPL). Se basa en support vector machines, incorpora modelos para español, catalán e inglés.
- TreeTagger
- Original Xerox Tagger
- μ-TBL
- QTA
- Lingua-EN-Tagger
- PoSTech Korean morphological analyzer and tagger

analizadores morfológicos

- Stanford POS tagger
- Brill's Transformation-based learning Tagger
- Maximum Entropy part of speech tagger MXPOST
- TnT
- SVMTool
- TreeTagger
- Original Xerox Tagger Common lisp, basado en HMM. También hay una versión para el español.
- μ-TBL
- QTA
- Lingua-EN-Tagger
- PoSTech Korean morphological analyzer and tagger

analizadores morfológicos

- Stanford POS tagger
- Brill's Transformation-based learning Tagger
- Maximum Entropy part of speech tagger MXPOST
- TnT
- SVMTool
- TreeTagger
- Original Xerox Tagger
- μ-TBL
- QTA
- Lingua-EN-Tagger
- PoSTech Korean morphological analyzer and tagger

analizadores morfológicos

- Stanford POS tagger
- Brill's Transformation-based learning Tagger
- Maximum Entropy part of speech tagger MXPOST
- TnT
- SVMTool
- TreeTagger
- Original Xerox Tagger
- μ-TBL Prolog, aprendizaje basado en transformaciones, también se puede usar para otro tipo de aprendizaje.
- QTA
- Lingua-EN-Tagger
- PoSTech Korean morphological analyzer and tagger

analizadores morfológicos

- Stanford POS tagger
- Brill's Transformation-based learning Tagger
- Maximum Entropy part of speech tagger MXPOST
- TnT
- SVMTool
- TreeTagger
- Original Xerox Tagger
- μ-TBL
- QTA java (Archivos de clases, no fuente). Basado en HMM. Incluye diccionarios del inglés y del alemán.
- Lingua-EN-Tagger
- PoSTech Korean morphological analyzer and tagger

analizadores morfológicos

- Stanford POS tagger
- Brill's Transformation-based learning Tagger
- Maximum Entropy part of speech tagger MXPOST
- TnT
- SVMTool
- TreeTagger
- Original Xerox Tagger
- μ-TBL
- QTA
- Lingua-EN-Tagger
- PoSTech Korean morphological analyzer and tagger

analizadores morfológicos

- Stanford POS tagger
- Brill's Transformation-based learning Tagger
- Maximum Entropy part of speech tagger MXPOST
- TnT
- SVMTool
- TreeTagger con diccionarios para inglés, alemán, francés, italiano, ruso y creciendo. Para Solaris y Linux. Basado en árboles de decisión.
- Original Xerox Tagger
- μ-TBL
- QTA
- Lingua-EN-Tagger
- PoSTech Korean morphological analyzer and tagger

analizadores morfológicos

- Stanford POS tagger
- Brill's Transformation-based learning Tagger
- Maximum Entropy part of speech tagger MXPOST
- TnT
- SVMTool
- TreeTagger
- Original Xerox Tagger
- μ-TBL
- QTA
- Lingua-EN-Tagger Perl, basado en HMM de 2 palabras.
- PoSTech Korean morphological analyzer and tagger

análisis de grupos lingüísticos

se identifican grupos lingüísticos o *chunks*: p.ej., *[el gato] [come] [pescado]*:

- YamCha
- LingPipe
- fnTBL

análisis de grupos lingüísticos

se identifican grupos lingüísticos o *chunks*: p.ej., *[el gato] [come] [pescado]*:

- YamCha C/C++ código abierto, para el inglés, ganador de un concurso en reconocimiento de entidades con nombre (p.ej.: *Woody Allen*)
- LingPipe
- fnTBL

análisis de grupos lingüísticos

se identifican grupos lingüísticos o *chunks*: p.ej., *[el gato] [come] [pescado]*:

- YamCha
- LingPipe java (GPL), reconoce entidades con nombre, finales de oración, e incluso co-referencia dentro de un documento
- fnTBL

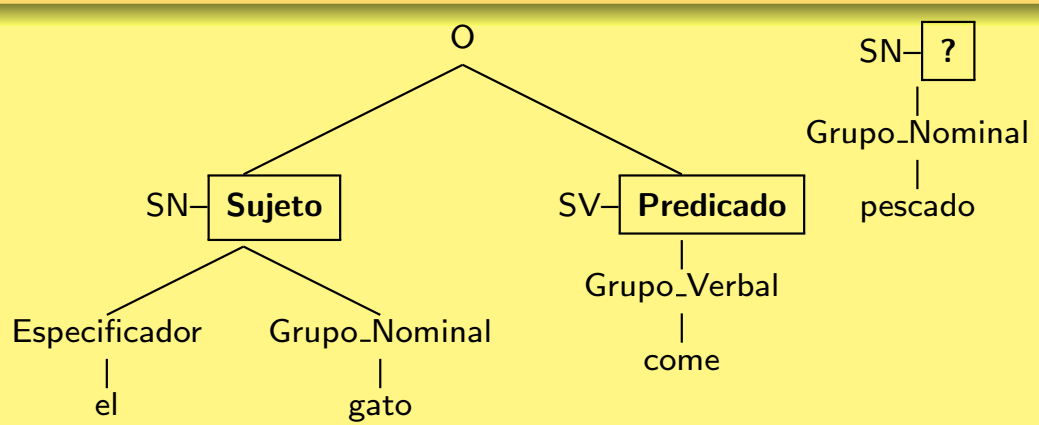
análisis sintáctico tradicional (manual)

1. uno (o más) lingüistas crean una gramática de la lengua
   - reglas independientes de contexto (*Context Free Grammar*)
     $SN \rightarrow Det\ N$
   - reglas enriquecidas con rasgos (*Unification Grammar*)
     $SN_{fem,sg} \rightarrow Det_{fem,sg}\ N_{fem,sg}$
   - basada en el léxico (*Lexicalized Grammar*)
     $SN_{gato} \rightarrow Det\ N_{gato}$
2. un analizador (o *parser*) utiliza esta gramática para asignar estructura a oraciones no vistas previamente

análisis sintáctico tradicional (manual)

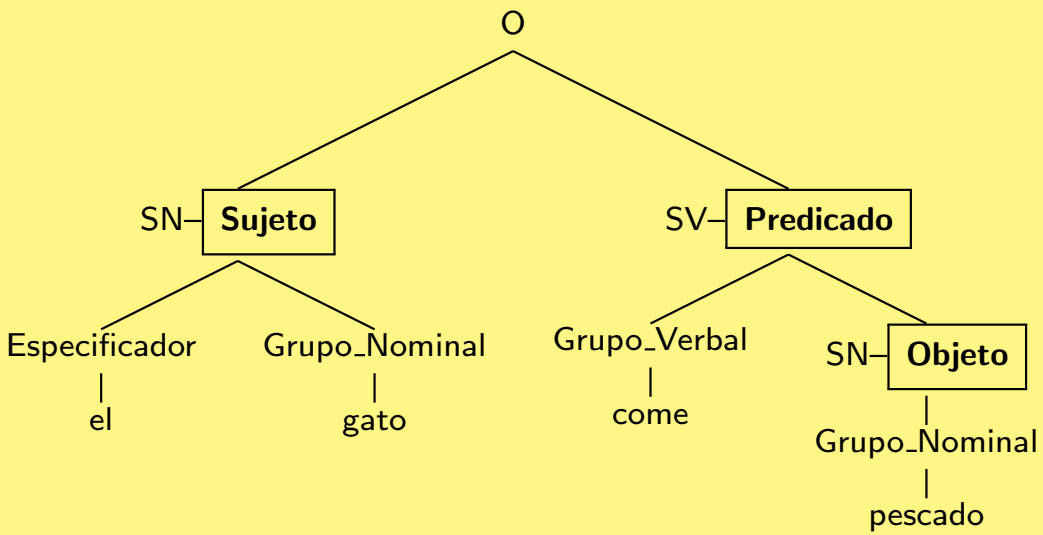
1. uno (o más) lingüistas crean una gramática de la lengua
2. un analizador (o *parser*) utiliza esta gramática para asignar estructura a oraciones no vistas previamente

SV → V Objeto



análisis sintáctico tradicional (manual)

1. uno (o más) lingüistas crean una gramática de la lengua
2. un analizador (o *parser*) utiliza esta gramática para asignar estructura a oraciones no vistas previamente

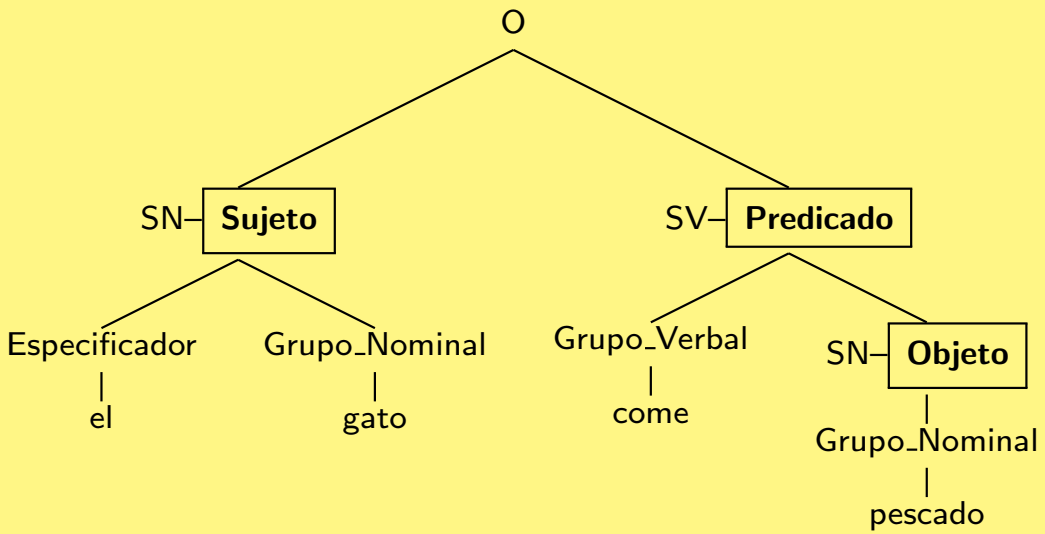


análisis sintáctico basado en gramáticas manuales

- Prolog tiene una extensión para implementar gramáticas libres de contexto: DCG (Definite Clause Grammars)
- ALE es un analizador para gramáticas de unificación basada en prolog, incluye gramáticas del inglés en HPSG (una clase famosa de gramáticas de unificación)
- Link Grammar C, código abierto, basada en formalismo de dependencias
- English Resource Grammar gramática HPSG del inglés, funciona sobre LKB
- Jacy gramática HPSG del japonés
- Modern Greek Resource Grammar gramática HPSG para el griego moderno

cómo se obtienen *parsers* probabilísticos

corpus de entrenamiento



cómo se obtienen *parsers* probabilísticos

corpus de entrenamiento

método de inferencia

gramáticas libres de contexto probabilísticas (lexicalizadas)
(Probabilistic (lexicalized) Context Free Grammars)

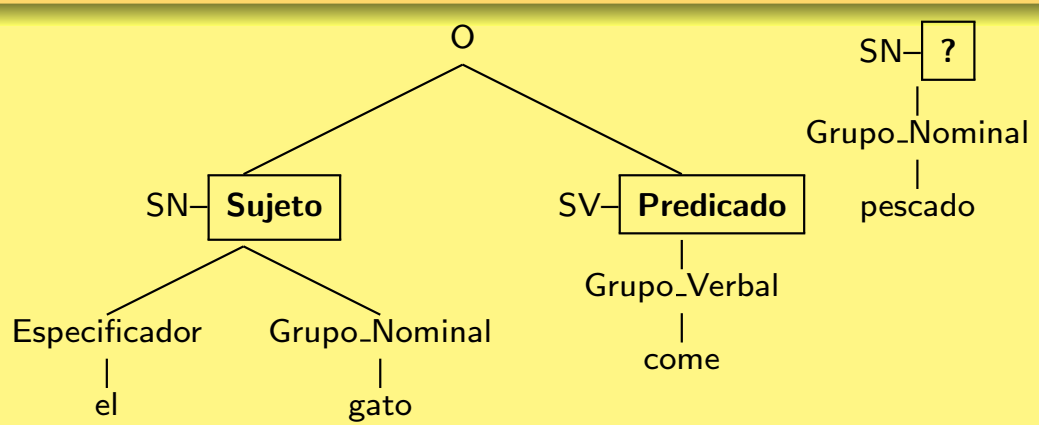
cómo se obtienen *parsers* probabilísticos

corpus de entrenamiento

método de inferencia

herramienta de análisis

SV $\rightarrow$ V Objeto $P = .82$
SV $\rightarrow$ V Circunstancial $P = .18$

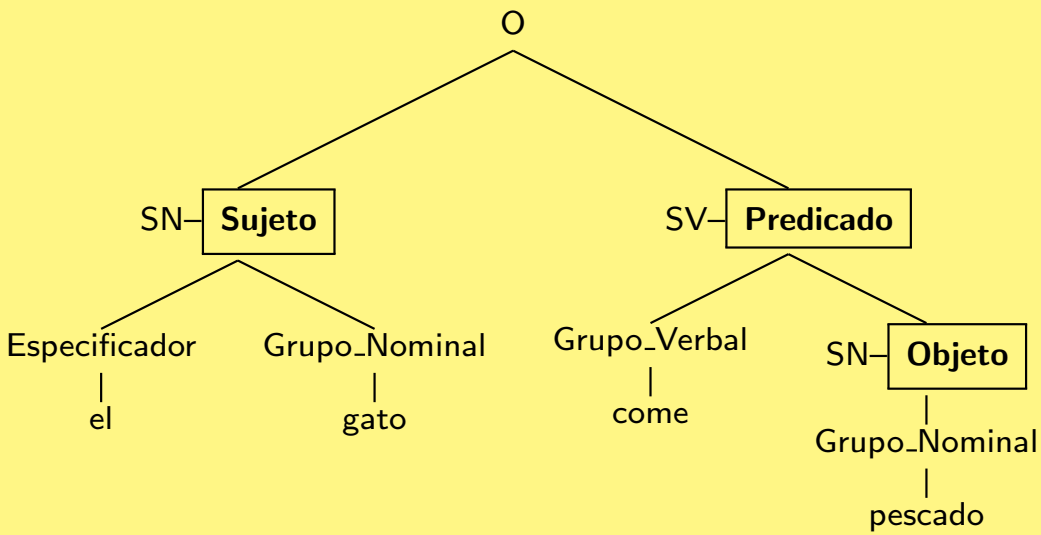


cómo se obtienen *parsers* probabilísticos

corpus de entrenamiento

método de inferencia

herramienta de análisis



analizadores probabilísticos: corpus anotados

la mayor parte de corpus son pagos, excepto unos pocos, que son chicos :(

- AnCora_ES
- Susanne
- Christine
- Lucy

analizadores sintácticos probabilísticos

- software by Mark Jonhnson
- MINIPAR
- Stanford Lexicalized Parser
- Eugene Charniak's parser
- Michael Collins' parser
- Dan Bikel's parser
- Apple Pie Parser
- Malt Parser
- DeSR Parser

analizadores sintácticos probabilísticos

- software by Mark Jonhnson Mark Johnson tiene disponible en su página web un montón de software relacionado con parsing, incluyendo un parser basado en reranking del 2005, una implementación en C muy eficiente de un parser clásico (CKY) y un parser muy popular en common lisp
- MINIPAR
- Stanford Lexicalized Parser
- Eugene Charniak's parser
- Michael Collins' parser
- Dan Bikel's parser
- Apple Pie Parser
- Malt Parser
- DeSR Parser

analizadores sintácticos probabilísticos

- software by Mark Jonhnson
- MINIPAR C++, código abierto, para el inglés, muy eficiente y muy claro
- Stanford Lexicalized Parser
- Eugene Charniak's parser
- Michael Collins' parser
- Dan Bikel's parser
- Apple Pie Parser
- Malt Parser
- DeSR Parser

analizadores sintácticos probabilísticos

- software by Mark Jonhnson
- MINIPAR
- Stanford Lexicalized Parser java, código abierto, para el inglés
- Eugene Charniak's parser
- Michael Collins' parser
- Dan Bikel's parser
- Apple Pie Parser
- Malt Parser
- DeSR Parser

analizadores sintácticos probabilísticos

- software by Mark Jonhnson
- MINIPAR
- Stanford Lexicalized Parser
- Eugene Charniak's parser C++, código abierto, para el inglés
- Michael Collins' parser
- Dan Bikel's parser
- Apple Pie Parser
- Malt Parser
- DeSR Parser

analizadores sintácticos probabilísticos

- software by Mark Jonhnson
- MINIPAR
- Stanford Lexicalized Parser
- Eugene Charniak's parser
- Michael Collins' parser C, fuente y ejecutables, para el inglés, también existe una versión que se puede correr como un daemon, documentación de su adaptación al checo
- Dan Bikel's parser
- Apple Pie Parser
- Malt Parser
- DeSR Parser

analizadores sintácticos probabilísticos

- software by Mark Jonhnson
- MINIPAR
- Stanford Lexicalized Parser
- Eugene Charniak's parser
- Michael Collins' parser
- Dan Bikel's parser java, código abierto y clases, incluye una reimplementación exacta del parser de Collins y packs para inglés, chino y árabe, e está trabajando en adaptaciones al español y al coreano
- Apple Pie Parser
- Malt Parser
- DeSR Parser

qué entendemos por análisis semántico?

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

para ello hay que asociar cada palabra a un *sentido*

diccionarios de sentidos y ontologías

Existen diversos diccionarios de sentidos organizados en forma de árbol (*ontologías léxicas*):

- WordNet
- EuroWordNet
- MCR
- KnowNet

diccionarios de sentidos y ontologías

Existen diversos diccionarios de sentidos organizados en forma de árbol (*ontologías léxicas*):

- WordNet 155.00 nombres, verbos y adjetivos del inglés se organizan en grupos de sinónimos (*synsets*) que a su vez se relacionan entre ellos mediante relaciones semánticas: *tipo de*, *contrario de*, etc. Totalmente libre, en varios formatos de uso y consulta y con extensa documentación, científica y técnica.
- EuroWordNet
- MCR
- KnowNet

diccionarios de sentidos y ontologías

Existen diversos diccionarios de sentidos organizados en forma de árbol (*ontologías léxicas*):

- WordNet
- EuroWordNet usando la estructura de WordNet como esqueleto común (Inter-Lingual-Index, ILI) se construyen ontologías para español, holandés, italiano, alemán, francés, checo y estonio. Libres para uso no comercial
- MCR
- KnowNet

diccionarios de sentidos y ontologías

Existen diversos diccionarios de sentidos organizados en forma de árbol (*ontologías léxicas*):

- WordNet
- EuroWordNet
- MCR Multilingual Central Repository, integra en el marco de EuroWordnet framework, a través del Interlingual Index:
  - five local wordnets and six versions of English WordNet,
  - WordNet Domains (Magnini and Cavaglià 2000),
  - new versions of the Base Concepts and the Top Concept Ontology (Álvez et al. 2008),
  - the SUMO ontology (Niles and Pease 2001),
  - and hundreds of thousands of automatically acquired semantic relations (MCR has 934.771 vs. 235.402 unique semantic relations in WordNet 3.0).
- KnowNet

diccionarios de sentidos y ontologías

Existen diversos diccionarios de sentidos organizados en forma de árbol (*ontologías léxicas*):

- WordNet
- EuroWordNet
- MCR
- KnowNet an extension of WordNet where topical relations between synsets are added. It is automatically created by semantically disambiguating small portions of Topic Signatures acquired from the web (Martínez et al. 08), then connecting large sets of topically-related concepts.

desambiguación de sentidos

es un problema difícil, y en los últimos años se orienta más hacia el **descubrimiento de sentidos**

- concursos Senseval
- concursos SemEval
- algoritmos para la asignación de palabras a sentidos basados en WordNet

FreeLing tiene integrados algunos algoritmos de desambiguación, incluyendo un método para seleccionar Basic Level Concepts de WordNet (Basic Level Concepts (BLC) are those concepts that are frequent and salient; they are neither overly general nor too specific).

otros recursos con información léxico-semántica

- SENSEM verbal data base and lexicon
- ADESSE
- Spanish FrameNet
- Unified Verb Index is a system which merges links and web pages from four different natural language processing projects
- Beth Levin's verbal classes
- PropBank
- FrameNet
- VerbNet

recuperación de información (*information retrieval*)

- Search Tools
- IN TeraScale Retrieval
- REtrieval COmponent INtegrator
- The Lemur Toolkit

recuperación de información (*information retrieval*)

- Search Tools un directorio que ayuda a encontrar el motor de búsqueda (*search engine*) más adecuado para cada necesidad: para web, intranets, diferentes tipos de datos, de aplicación, etc., con un apartado especial para motores de código abierto, incluyendo un artículo comparativo.
- IN TeraScale Retrieval
- REtrieval COmponent INtegrator
- The Lemur Toolkit

recuperación de información (*information retrieval*)

- Search Tools
- IN TeraScale Retrieval C++, GNU, un toolkit completo de herramientas de IR para todos los sistemas POSIX, con émfasis en recuperación de información semiestructurada (HTML, XML)
- REtrieval COmponent INtegrator
- The Lemur Toolkit

recuperación de información (*information retrieval*)

- Search Tools
- IN TeraScale Retrieval
- REtrieval COmponent INtegrator herramientas libres para desarrollar investigación en recuperación de información
- The Lemur Toolkit

recuperación de información (*information retrieval*)

- Search Tools
- IN TeraScale Retrieval
- REtrieval COmponent INtegrator
- The Lemur Toolkit explota el trabajo de modelado de lenguaje en otras áreas de PLN para aplicarlo a recuperación de información, orientado sobretodo a investigación

traducción automática (*machine translation*)

- Apertium
- Delph-In
- Laurie's links
- la serie de workshops sobre Teaching Machine Translation (con interesantes artículos sobre recursos libres): 2001, 2003

traducción automática (*machine translation*)

- Apertium un traductor entre lenguas romances de España, código abierto, basado en análisis superficial dentro del proyecto OpenTrad, que también desarrolla un traductor de código abierto entre castellano y euskera, basado en análisis sintáctico completo
- Delph-In
- Laurie's links
- la serie de workshops sobre Teaching Machine Translation (con interesantes artículos sobre recursos libres): 2001, 2003

traducción automática (*machine translation*)

- Apertium
- Delph-In es un proyecto de comprensión profunda de lenguaje natural cuyos recursos (libres!) han sido aplicados a traducción automática
- Laurie's links
- la serie de workshops sobre Teaching Machine Translation (con interesantes artículos sobre recursos libres): 2001, 2003

traducción automática (*machine translation*)

- Apertium
- Delph-In
- Laurie's links una exhaustiva lista de motores de traducción disponibles via web (en el año 2000), donde se especifica los idiomas que tratan, el texto máximo permitido, etc.
- la serie de workshops sobre Teaching Machine Translation (con interesantes artículos sobre recursos libres): 2001, 2003

traducción automática estadística y corpus alineados

La idea básica de los sistemas de traducción automática estadística es obtener un diccionario bilingüe a partir de corpus paralelos en las dos lenguas, que han sido alineados.

traducción automática estadística y corpus alineados

La idea básica de los sistemas de traducción automática estadística es obtener un diccionario bilingüe a partir de corpus paralelos en las dos lenguas, que han sido alineados.

el	*the*
gato	*cat*
come	*eats*
pescado	*fish*

traducción automática estadística y corpus alineados

La idea básica de los sistemas de traducción automática estadística es obtener un diccionario bilingüe a partir de corpus paralelos en las dos lenguas, que han sido alineados.

- Europarl corpus de documentos de la Unión Europea, con cerca de 20 millones de palabras en total, con unas 740.000 frases de cada una de las 11 lenguas, alineados manualmente a nivel de oración.
- Hansards
- CRATER
- OPUS
- GNOME's GUI messages translation statistics
- Emille

traducción automática estadística y corpus alineados

La idea básica de los sistemas de traducción automática estadística es obtener un diccionario bilingüe a partir de corpus paralelos en las dos lenguas, que han sido alineados.

- Europarl
- Hansards corpus de documentos del parlamento de Canadá, paralelos en inglés y francés, alineados a nivel de oración o menor
- CRATER
- OPUS
- GNOME's GUI messages translation statistics
- Emille

traducción automática estadística y corpus alineados

La idea básica de los sistemas de traducción automática estadística es obtener un diccionario bilingüe a partir de corpus paralelos en las dos lenguas, que han sido alineados.

- Europarl
- Hansards
- CRATER corpus alineado trilingüe: inglés, francés y castellano, con anotaciones morfosintácticas revisadas manualmente.
- OPUS
- GNOME's GUI messages translation statistics
- Emille

traducción automática estadística y corpus alineados

La idea básica de los sistemas de traducción automática estadística es obtener un diccionario bilingüe a partir de corpus paralelos en las dos lenguas, que han sido alineados.

- Europarl
- Hansards
- CRATER
- OPUS corpus de manuales técnicos (e.g., software libre, documentación de la Unión Europea) traducidos a varios idiomas, que han sido alineados automáticamente, están disponibles el corpus y el software de alineación.
- GNOME's GUI messages translation statistics
- Emille

traducción automática estadística y corpus alineados

La idea básica de los sistemas de traducción automática estadística es obtener un diccionario bilingüe a partir de corpus paralelos en las dos lenguas, que han sido alineados.

- Europarl
- Hansards
- CRATER
- OPUS
- GNOME's GUI messages translation statistics
- Emille corpus de 200.000 palabras, paralelo en inglés, hindi, bengalí, punjabí, gujarati y urdu.

traducción automática estadística y corpus alineados

La idea básica de los sistemas de traducción automática estadística es obtener un diccionario bilingüe a partir de corpus paralelos en las dos lenguas, que han sido alineados.

- Moses a statistical machine translation system that allows you to automatically train translation models for any language pair. All you need is a collection of translated texts (parallel corpus).
- EGYPT es un toolkit para desarrollar sistemas de traducción automática estadística a partir de corpus paralelos.
- Rada Mihalcea mantiene un extenso repositorio sobre alineación para traducción automática: corpus, software, etc.

exploración de datos: clasificación y clustering

- R la versión libre de S: un software para computación estadística y gráficos. Para todas las plataformas.
- The 'Bow' Toolkit librerías en C para análisis estadístico de texto, modelado de lenguaje y recuperación de información.
- Weka algoritmos para todo tipo de tareas de minería de datos, se pueden usar desde terminal, desde interfaz gráfica o desde tu propio código java. Cuenta con un libro de introducción a Weka y a la minería de datos en general y una activa lista de usuarios.
- Mallet es una herramienta para aplicar todo tipo de técnicas de Machine Learning a lenguaje natural

entornos para crear analizadores

- NLTK - Natural Language Toolkit
- GATE - a General Architecture for Text Engineering
- CCG Library
- EDG
- LKB
- Matrix
- NLPfarm
- Ellogon

entornos para crear analizadores

- NLTK - Natural Language Toolkit una suite de librerías y programas en Python para desarrollar gramáticas y analizadores de todo tipo, simbólico y estadístico
- GATE - a General Architecture for Text Engineering
- CCG Library
- EDG
- LKB
- Matrix
- NLPfarm
- Ellogon

entornos para crear analizadores

- NLTK - Natural Language Toolkit
- GATE - a General Architecture for Text Engineering java, código abierto, muy bien documentado, resultado de un gran proyecto, ampliamente usado para diversas tareas de PLN, sobretodo orientado a comprensión profunda
- CCG Library
- EDG
- LKB
- Matrix
- NLPfarm
- Ellogon

entornos para crear analizadores

- NLTK - Natural Language Toolkit
- GATE - a General Architecture for Text Engineering
- CCG Library una colección de herramientas para desarrollar analizadores en el marco de Combinatory Categorial Grammar, java, código abierto, LGPL o librería GNU
- EDG
- LKB
- Matrix
- NLPfarm
- Ellogon

entornos para crear analizadores

- NLTK - Natural Language Toolkit
- GATE - a General Architecture for Text Engineering
- CCG Library
- EDG Example-based Development of Grammars, un sistema en lisp para desarrollar analizadores en el marco de Head Driven Phrase Structure Grammar (HPSG)
- LKB
- Matrix
- NLPfarm
- Ellogon

entornos para crear analizadores

- NLTK - Natural Language Toolkit
- GATE - a General Architecture for Text Engineering
- CCG Library
- EDG
- LKB entorno para desarrollar gramáticas y léxicos basados en gramáticas de unificación, explotando los principios de estructuras tipadas del proyecto DELPH-IN
- Matrix
- NLPfarm
- Ellogon

entornos para crear analizadores

- NLTK - Natural Language Toolkit
- GATE - a General Architecture for Text Engineering
- CCG Library
- EDG
- LKB
- Matrix un kit de principiante para desarrollar gramáticas HPSG en LKB
- NLPfarm
- Ellogon

entornos para crear analizadores

- NLTK - Natural Language Toolkit
- GATE - a General Architecture for Text Engineering
- CCG Library
- EDG
- LKB
- Matrix
- NLPfarm concentra diversos módulos para procesamiento de diálogo en java
- Ellogon

entornos para crear analizadores

- NLTK - Natural Language Toolkit
- GATE - a General Architecture for Text Engineering
- CCG Library
- EDG
- LKB
- Matrix
- NLPfarm
- Ellogon entorno gráfico multiplataforma para todo tipo de aplicaciones de ingeniería del lenguaje

entornos para crear corpus anotados

- Alembic un banco de trabajo (*workbench*) para desarrollar corpus anotados y analizadores que se basen en ellos con una gran reducción del esfuerzo humano
- Wordfreak una herramienta de anotación java (mozilla public license 1.1), para anotaciones humanas, automáticas y semiautomáticas (mediante active learning)
- AGTK herramienta para anotar señales acústicas y todo tipo de series temporales (audio, video), basada en grafos

directorios de herramientas, recursos y documentación

- OpenNLP
- el grupo de PLN de Stanford
- Kenji Kita
- Manuel Barberá
- recursos del Summer Institute of Linguistics
- recursos de la Linguist List
- WEBIR/IE

directorios de herramientas, recursos y documentación

- OpenNLP es un directorio de recursos de PLN de código abierto en sourceforge
- el grupo de PLN de Stanford
- Kenji Kita
- Manuel Barberá
- recursos del Summer Institute of Linguistics
- recursos de la Linguist List
- WEBIR/IE

directorios de herramientas, recursos y documentación

- OpenNLP
- el grupo de PLN de Stanford mantiene lista de recursos y herramientas de PLN probabilísticas y de lingüística computacional muy extenso y actualizado
- Kenji Kita
- Manuel Barberá
- recursos del Summer Institute of Linguistics
- recursos de la Linguist List
- WEBIR/IE

directorios de herramientas, recursos y documentación

- OpenNLP
- el grupo de PLN de Stanford
- Kenji Kita también tiene una extensa página de links a recursos y herramientas para PLN
- Manuel Barberá
- recursos del Summer Institute of Linguistics
- recursos de la Linguist List
- WEBIR/IE

directorios de herramientas, recursos y documentación

- OpenNLP
- el grupo de PLN de Stanford
- Kenji Kita
- Manuel Barberá también mantiene una muy respetable lista de enlaces, poco actualizados pero con el interés de centrarse bastante en lenguas romances
- recursos del Summer Institute of Linguistics
- recursos de la Linguist List
- WEBIR/IE

directorios de herramientas, recursos y documentación

- OpenNLP
- el grupo de PLN de Stanford
- Kenji Kita
- Manuel Barberá
- recursos del Summer Institute of Linguistics orientados sobretodo a la descripción de lenguas
- recursos de la Linguist List
- WEBIR/IE

directorios de herramientas, recursos y documentación

- OpenNLP
- el grupo de PLN de Stanford
- Kenji Kita
- Manuel Barberá
- recursos del Summer Institute of Linguistics
- recursos de la Linguist List cubren todo el espectro de la lingüística: descripción, aprendizaje, diccionarios, fonética, lingüística histórica... y por supuesto PLN
- WEBIR/IE

directorios de herramientas, recursos y documentación

- OpenNLP
- el grupo de PLN de Stanford
- Kenji Kita
- Manuel Barberá
- recursos del Summer Institute of Linguistics
- recursos de la Linguist List
- WEBIR/IE recursos de IR, publicaciones, conferencias, contactos, listas de noticias...

instituciones

- ELDA - ELRA Evaluations and Language resources Distribution Agency, tiene un completísimo catálogo de recursos lingüísticos para lenguas europeas, libres y pagos.
- HLT central Human Language Technology, repositorio europeo de grupos y entidades relacionados con las tecnologías del lenguaje, tiene un completo calendario de eventos y enlaces interesantes
- ELSNET European Network of Excellence in Human Language Technologies, con calendario de eventos (es uno de los principales sponsors del área), grupos relacionados, asociaciones, bolsa de trabajo y de becas, etc.
- ACL The Association for Computational Linguistics tiene enlaces a las principales conferencias mundiales sobre lenguaje natural, a la mayor parte de artículos que se publican en el área, a recursos de carácter general y a grupos de interés en

empresas

- MITRE tiene muchos proyectos de investigación en lenguaje natural, muchos con recursos libres
- Xerox
- AT&T

tarea para el hogar

1. buscar corpus (1000 palabras) en dos lenguas distintas del inglés para las que exista un tagger disponible libremente (preferiblemente freeling)
2. enviar los textos al grupo de google
3. pensar un detector de idioma

tarea extra para el hogar

- familiarizarse con el formato de los archivos de entrada para weka (.arff)
- crear un script para convertir textos en vectores...

hasta mañana!