Universidad
Carlos III de Madrid

Módulo IX
Extracción de Información

OpenCourseWare

Recuperación y Acceso a la Información

Contenidos

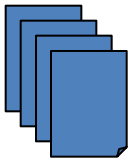
Extracción de información

Reconocimiento de Entidades de Nombre (*Named Entities)*

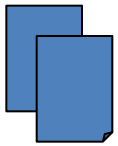
- Competiciones
- Clasificaciones
- Aplicaciones
- Atributos
- Técnicas
- Problemas

Extracción de Información vs Recuperación de Información

RI



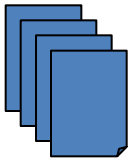
consulta



Corpus

Documentos ¿con la respuesta?

EI



consulta

#Element1
#Element2
#Element3
...

Corpus

Sucesos, entidades o relaciones ¿válidos?

Extracción de Información. Definición

Proceso de seleccionar, clasificar y combinar datos que están presentes explícitamente (no implícita) en uno o más documentos en lenguaje natural.

Information extraction is the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, making the information more suitable for information processing tasks.

Marie-Francine Moens: "Information Extraction: Algorithms and Prospects in a Retrieval Context". Springer, 2006

Contenidos

Extracción de información

Reconocimiento de Entidades de Nombre (*Named Entities)*

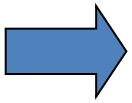
- Competiciones
- Clasificaciones
- Aplicaciones
- Atributos
- Técnicas
- Problemas

Extracción de Información. Tareas

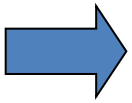
- Reconocimiento de la Entidad del Nombre (personas, sitios, organismos,…)
  - **John Smith** works for **IBM**
- Resolución de correferencia nominal
  - Bill Clinton went to New York where **he** was invited. The former…
- Reconocimiento del rol semántico
  - She clapped (**agent**) her hands (**body part**) in inspiration (**cause**)
- Reconocimiento de relaciones entre entidades (normalmente tripletas como Fundación(UC3M, 1991))
  - John Smith **works for** IBM -> Relation works for
- Timex y reconocimiento de la línea temporal
  - On **April 16, 2005** I passed the exam and I had studied a lot three weeks before
- Plantillas de escenarios: formularios rellenados con entidades, puede implicar más de un documento

Ejemplo de Extracción de Información

<DOC>
<DOCID> wsj93_050.0203 </DOCID>
<DOCNO> 930219-0013. </DOCNO>
<HL> Marketing Brief:@ Noted.... </HL>
<DD> 02/19/93 </DD>
<SO> WALL STREET JOURNAL (J), PAGE B5 </SO>
<CO> NYTA </CO>
<IN> MEDIA (MED), PUBLISHING (PUB) </IN>
<TXT> New York Times Co. named Russell T. Lewis, 45, president andgeneral manager of its flagship New York Times newspaper,responsible for all business-side activities. He was executivevice president and deputy general manager. He succeeds LanceR. Primis, who in September was named president and chiefoperating officer of the parent. </TXT>
</DOC>



<ORGANIZATION-1> "New York Times Co."
<ORGANIZATION-2> "New York Times"
<PERSON-1> "Russell T. Lewis"
<PERSON-2> "Lance R. Primis"



<SUCCESSION-1> ORGANIZATION :
<ORGANIZATION-2> POST : "president"
WHO_IS_IN : <PERSON-1>
WHO_IS_OUT : <PERSON-2>

Reconocimiento de Entidades de Nombre

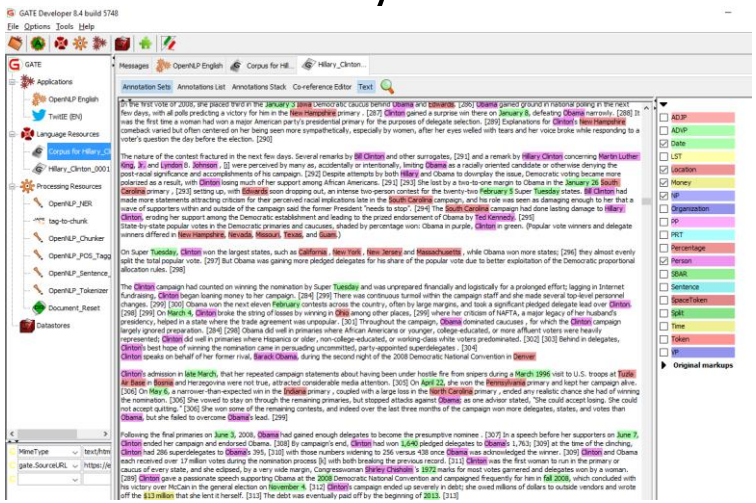
- Pueden ser unidades simples o unidades multipalabra (CC.OO., CCOO, J. A. Borbón, 18:08, 1 de abril, 6036-BJF, M-1967-JM, un millón de euros, Dr. Borbón)
- Suelen ser identificadores únicos de entidades (similar a instancias)
- Pertenecen a una clasificación semántica. Hay categorías reconocidas en dominios generales (persona, organización, lugar, etc.) y específicos (proteínas, genes, etc.).
- Generalmente responden a términos específicos y relevantes de dominio

- Demos:
  - http://cogcomp.cs.illinois.edu/demo/ner/results.php
  - http://nlp.stanford.edu:8080/ner/
  - http://alias-i.com/lingpipe/web/demo-ne.html
  - http://nlp.lsi.upc.edu/freeling/demo/demo.php
  - http://nactem7.mib.man.ac.uk/geniatagger/

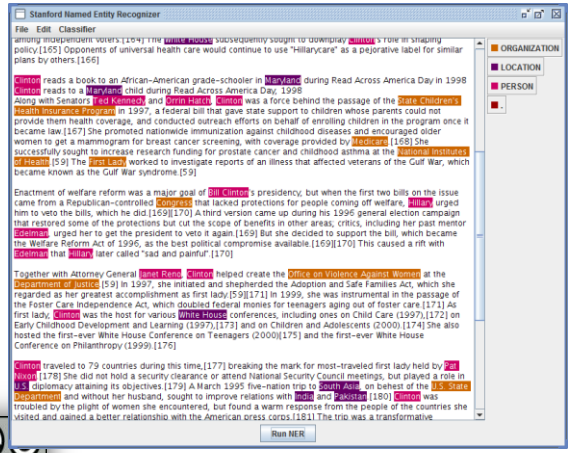
  Varios: http://www.clt.gu.se/wiki/interactive-online-demos
  - https://www.meaningcloud.com/es/demo

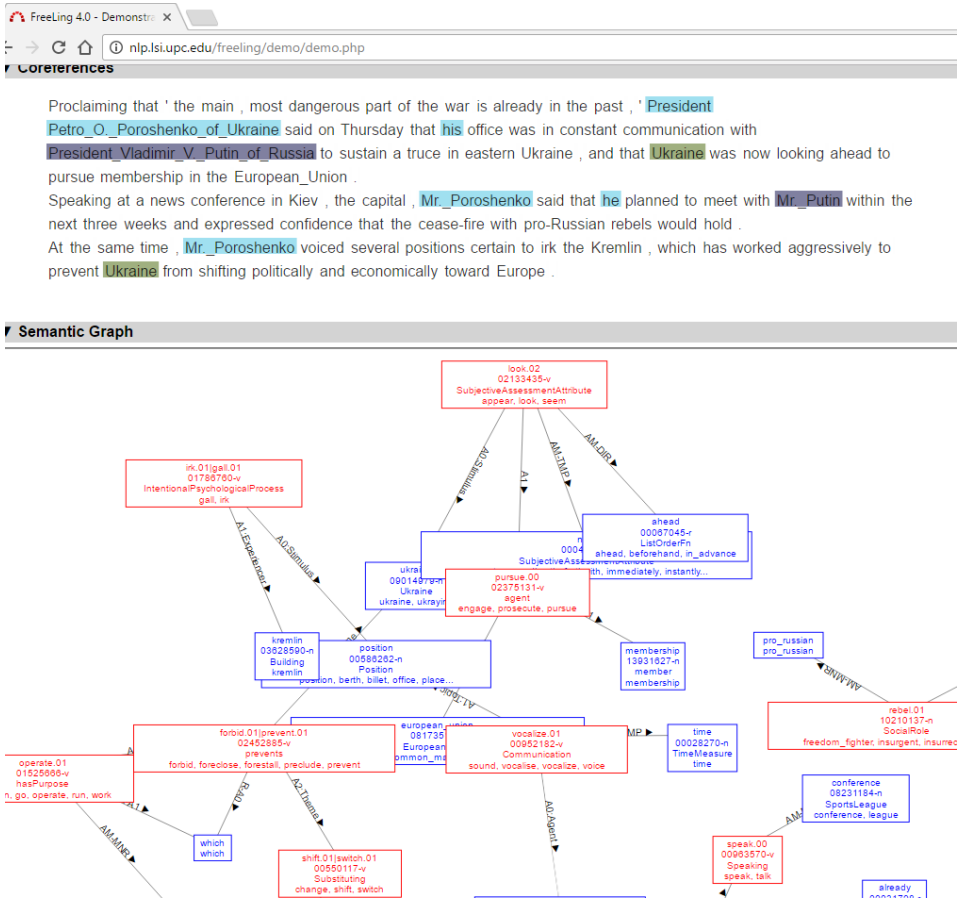
Reconocimiento de Entidades de Nombre (II)

- Con Gate y OPENNLP



- Con Stanford NER





- Con Freeling

Competiciones y conferencias

- Message Understanding Conference (MUC). Fueron iniciadas en 1987 financiadas por DARPA. Su principal objetivo era evaluar y promover el desarrollo de la extracción de información. Se celebraron siete ediciones hasta 1999.
  http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

- International Conference on Language Resources and Evaluation (LREC)

- Computational Natural Language Learning (CoNLL) workshop
  http://www.cnts.ua.ac.be/conll/

- Automatic Content Extraction (ACE), organizado por NIST
  http://www.nist.gov/speech/tests/ace/

Message Understanding Conferences

CONFER.	*AÑO*	*FUENTE*	*TOPIC*
MUC-1	1987	Mil. reports	Fleet Operations
MUC-2	1989	Mil. reports	Fleet Operations
MUC-3	1991	News reports	Terrorist activities in Latin America
MUC-4	1992	News reports	Terrorist activities in Latin America
MUC-5	1993	News reports	Corporate Joint Ventures, Microelectronic production
MUC-6	1995	News reports	Negotiation of Labor Disputes and Corporate management Succession
MUC-7	1997	News reports	Airplane crashes, and Rocket/Missile Launches

Se trataba de localizar causa, agente, tiempo y lugar del evento, consecuencias, etc.

Competiciones. Entidades reconocidas

- **MUC-6:**
  - Enamex: personas, organismos y localizaciones
  - Numex: cantidades numéricas (moneda, porcentajes)
  - Timex: fechas (date/time)
- **CONLL-02/03:** personas, organismos, localizaciones y miscelánea
- **ACE:** instalación (*facility*), entidad geo-politica, organización, localización, persona, vehículo, arma
  - Múltiples subtipos
  - Correferencia
  - Otras tareas: cantidades, expresiones temporales, relaciones y eventos.

"U.S. Fish and Wildlife Service"
<ENAMEX TYPE="ORGANIZATION">
U.S. Fish and Wildlife Service
</ENAMEX>
"North and South America"
<ENAMEX TYPE="LOCATION">
North
</ENAMEX> and
<ENAMEX TYPE="LOCATION">
South America
</ENAMEX>
DATE: expresión de tiempo completa o parcial
TIME: expresión de tiempo del día completa o parcial
"all of 1987"
<TIMEX TYPE="DATE" ALT="1987">
all of 1987
</TIMEX>

Entidades reconocidas en la ACE entity task.

Type	Subtypes
FAC (Facility)	Airport, building-grounds, path, plant, subarea-facility
GOE (Geo-political entity)	Continent, County-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province
LOC (localization)	Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body
ORG (Organization)	Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports
PER (Person)	Group, Indeterminate, Individual
VEH (Vehicle)	Air, Land, Subarea-Vehicle, Underspecified, Water
WEA (Weapon)	Biological, Blunt, Chemical, Exploding, Nuclear, Projectile, Sharp, Shooting, Underspecified

Principales foros

- Desde el 96 hasta el 2008 se trabaja con un conjunto muy limitado de entidades

MUC-7	CoNLL-03	ACE-08
Person	Person	Person (animal names excluded)
Organization	Organization	Organization
Localization	Localization	Localization / Geo-political name/ Facility
–	Miscellaneous	Person (as not structured group) / Vehicle / Weapon
Time / Date	–	Temporal expression (independent task)
Currency / Percentage	–	Quantity (independent task)

- Año tras año en la ACE se trabaja con los mismos tipos de entidades
- Los principales foros de evaluación trabajan en su mayoría con corpus anotados
  - Las guías de anotación dicen lo que es o no entidad. Pero la anotación es incoherente entre diferentes foros.

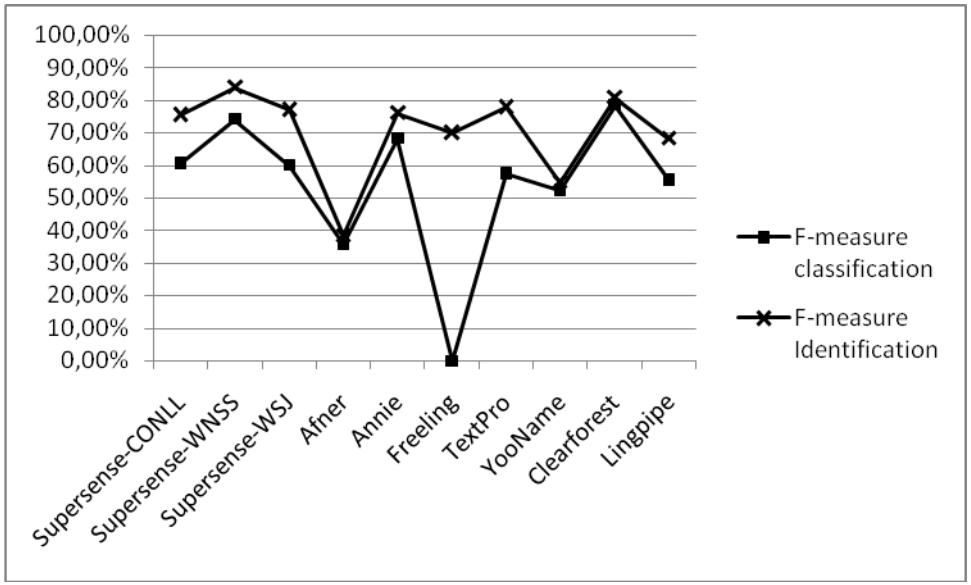
Phrase	CoNLL03	ACE	MUC-7	Disagreement
Baltimore defeated the Yankees	<Baltimore>ORG <Yankees>ORG	<Baltimore>ORG <Yankees>ORG	<Baltimore>LOC <Yankees>ORG	C
Zywiec Full Light	<Zywiec>ORG <Full Light>MISC	<Zywiec>ORG	<Zywiec>ORG	I, C
Empire State Building	<Empire State>LOC	<Empire State Building> FAC.Building	no markup	I, C, B
Alpine Skiing-Women 's World Cup Downhill	<World Cup>MISC	<Women>NOM <World>NOM	no markup	I, C, B
the new upper house of Czech parliament	<Czech>LOC	<Czech parliament>NOM	<parliament>ORG	I, C, B
Stalinist nations	<Stalinist>MISC	no markup	no markup	I
Wall Street Journal	<Wall Street Journal>ORG	<Wall Street Journal> ORG	no markup	I

Principales foros (II)

- La anotación llega a ser arbitraria en muchos casos
  - The guide itself establishes that in many situations it is confusing to differentiate between names and other mentions (e.g. “health ministry” is considered a nominal mention, but “ministry of health” is considered a name mention **due to the number of words following the so called “trumping rule”**).
- Foros más recientes relacionados con NER:
  - Desde 2007: XER (INEX Entity Ranking Track) – Ranking entidades
    - Ej. Input: I want a list of art galleries and museums in the Netherlands that have impressionist art – Output: articulos wikipedia
  - Desde 2009: KBP (Know. Base Population) en TAC (Text Analysis Conf.)
    - Ej. Input: “John Doe”, tipo “person”, [documento donde se cita], [listado de atributos] – Output: valores de atributos
  - Desde 2009: Entity Track de TREC
    - Ej. Aerolíneas que actualmente usan aviones Boeing 747 – páginas web aerolíneas ordenadas por relevancia

Evaluación de herramientas NER. Resultados (I)

- Resultados efectividad identificación y clasificación



- De las 10 herramientas, sólo tres superan el 70% en clasificación, sólo 2 superan el 80% en identificación

Evaluación de herramientas NER. Resultados (II)

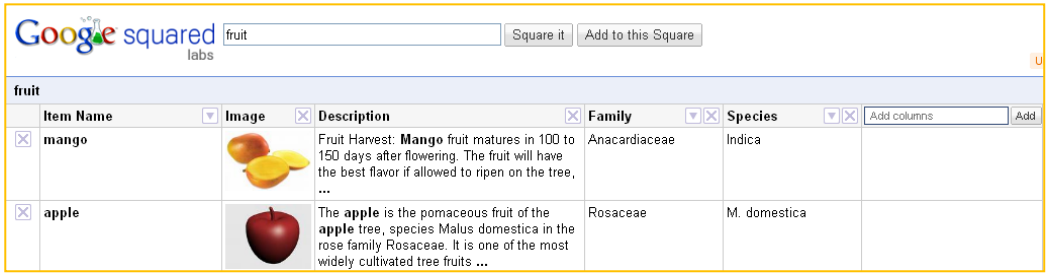
- Resultados por tipo de entidad: muy variables (ej. tipo persona: 30-70%), y no necesariamente relacionados con resultados globales

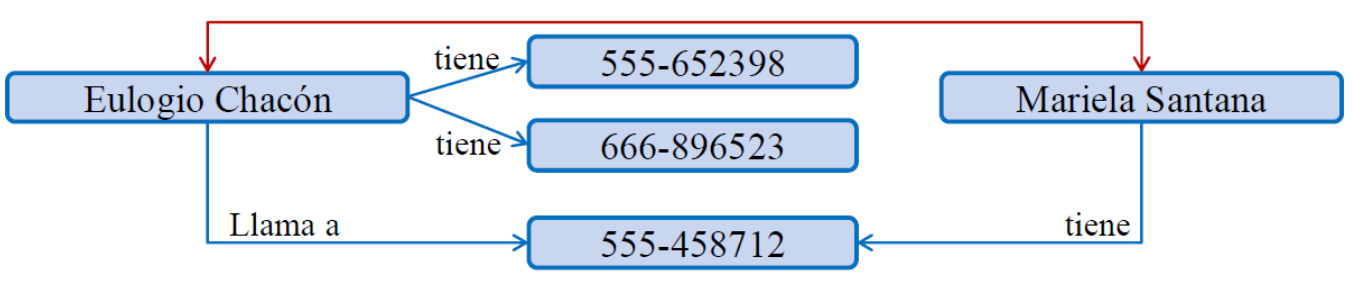
Tool	Entity type	N	F
Supers. CONLL	Person	32	0'63
	Location	43	0'64
	Org.	13	0'72
	Miscelanea	4	0
Supersense-WNSS	Person	38	0'65
	Location	48	0'78
	Group	25	0'88
	Time	9	0'66
	Quantity	9	0
	Food	6	1
	Communicat.	1	1
	Cognition	2	0'66
	Substance	1	0
	Relation	1	0
	Plant	6	1
	Object	1	1
	Other	1	1
Supersense-WSJ	Person	32	0'30
	Person-Desc.	5	1
	Geo-Pol.(other)	20	0'10
	Country	18	0'70
	State/Province	6	0'66
	Geo-Pol-Desc.	9	1
	Corporation	12	0'69
	Org.-Descrip.	12	0'86
	Date	10	1
	Money	4	0'28
	Food	7	1
	Ordinal	1	1
	Cardinal	7	0'92
Ling pipe	Person	32	0'67
	Location	47	0'47
	Org.	12	0'78

Tool	Entity type	N	F
YooName	Person	33	0'30
	Location	44	0'51
	Org.	13	0'88
	Vocation	4	1
	Country	21	0'66
	State/Prov.	4	0'75
	City	7	0'70
	Loc (other)	11	0
	Company	12	0'08
	Month	2	1
	Week Day	2	1
	Food	6	1
	Mineral	1	0
	Vegetal	1	1
Clear Forest	Person	53	0'72
	Country	19	0'97
	State/Prov.	6	1
	City	10	0'18
	Company	12	0'95
Text Pro	Person	32	0'59
	Location	44	0'51
	Org.	13	0'88
Afner	Person	32	0'50
	Location	44	0'36
	Org.	12	0
	Date	2	0'50
Annie	Person	32	73'3
	Person-title	2	1
	Location	44	62'5
	Organization	12	81'8
	Date	5	1
	Money	4	33'3
	Percent	2	1

Entidades de nombre. Aplicaciones

- Ayuda en la anotación gramatical de los textos
- Facilita la rápida comprensión de los textos
- Buscadores de instancias (por similitud o tipo)
  - Ej. Google Squared (Cancelado en 2011)
  - Actualmente distintos add-ons para Spreadsheets (la mayoría de pago, permiten 1000 consultas libres)





Entidades de nombre. Aplicaciones (II)

- Sistemas de búsqueda de respuestas (question-answering)
  - Identificación de respuestas concretas ante preguntas
    - Ej. Quién es el presidente de España?
    - Ej. Cuál es la capital de España?
  - Respuesta a algunas de las preguntas típicas del entorno periodístico, las 5 W: what, who, where, why, when, pero en dominios concretos a otras, p.e. en UMLS hay 54 tipos de relaciones
  - Problema añadido: temporalidad
- Enriquecimiento automático de ontologías, de gran importancia en la web semántica
- Rellenar automáticamente las bases de datos , como Yago, FreeBase, Dbpedia a partir de la Wikipedia
- Minería de opiniones (sentiment analysis)

Entidades de nombre. Aplicaciones (III)

Ejemplos de Minería de opiniones

http://sentistrength.wlv.ac.uk/

http://www.danielsoper.com/sentimentanalysis/default.aspx

Opal: Drupal plug-in for sentiment analysis

http://www.gi2mo.org/apps/opal-opinion-analyser/

OpenDover: análisis de comentarios en general

http://demo.opendover.nl/

http://java.opinionmining.nl/#

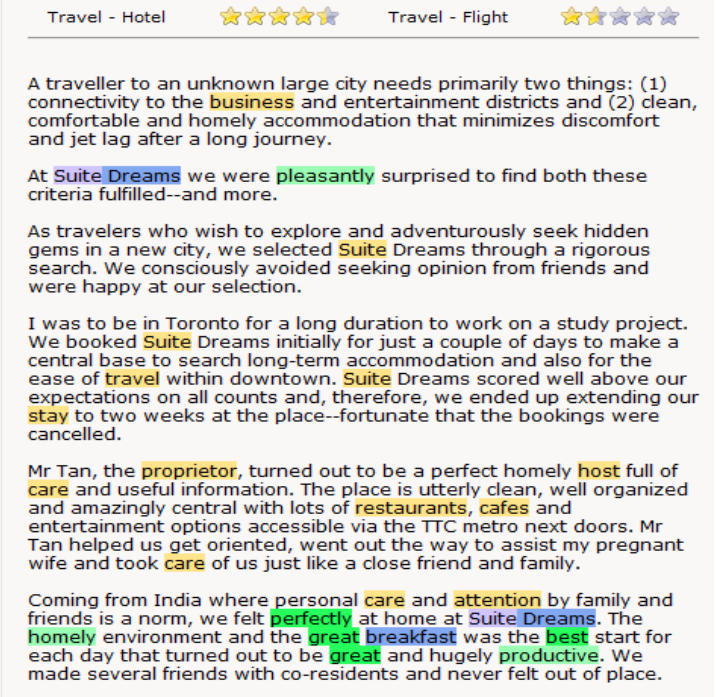
Análisis de twitter

http://www.socialmention.com/

http://hashtagify.me/

http://twittersentiment.appspot.com

http://www.pulseofthetweeters.com/



Acceder a www.tripadvisor.com y probar los 4 primeros ein http://demo.opendover.nl/

Atributos en el reconocimiento de entidades (I)

- Mayúsculas: si las palabras incorporan letras en mayúsculas, ya sea en la inicial, en su totalidad o alternativas.
- Tipo de caracteres: si las palabras contienen símbolos de puntuación como puntos, apóstrofes, guiones, comillas, etc., dígitos de cualquier tipo (cardinales, ordinales, caracteres romanos, etc.) o símbolos como la arroba, el ampersand, etc.
  - Depende de la tokenización
- Categoría morfo-sintáctica: categoría de la palabra (sustantivo, adjetivo, preposición, etc.), la forma normal de la misma, su lexema, raíz, y los prefijos y sufijos que pueda contener.
- La propia palabra, las palabras/tokens anteriores y/o posteriores (triggers)
- Longitud de la palabra

Atributos en el reconocimiento de entidades (II)

- La semántica de la palabra: su localización en recursos léxico-semánticos (listados → Gazeetters, tesauros, ontologías, etc. que me permiten determinar a qué tipo pertenece)

- Ocurrencia de otros elementos similares en el texto, sus características, localización en la sentencia o el documento, frecuencia y etiqueta asignada si ya ha sido procesada anteriormente.
  - Un ejemplo es la heurística aplicada por Nadeau (Nadeau 2007) en la que analiza si existen otras palabras iguales escritas en mayúsculas o minúsculas para determinar si se trata de una entidad o no.

- Meta-información: información asociada a la estructura del propio documento (por ejemplo, código HTML que contiene a un elemento, o sección XML) o referente a información general del mismo (por ejemplo, URI, cabecera, etc.).

Hearst (1991)

Hearst (1992)

- "Agar is a substance prepared from a mixture of **red algae, such as Gelidium,** for laboratory or industrial use"

Hearst para hiperónimos definió los siguientes indicadores:

- "Y such as X ((, X)* (, and|or) X)"
- "such Y as X"
- "X or other Y"
- "X and other Y"
- "Y including X"
- "Y, especially X"

Incluso patrones más complejos:

PERSON [be]? (named|appointed|*etc.*) Prep? ORG POSITION

- George Marshall was named US Secretary of State

Ejemplo extracción relaciones (UMLS)

Doppler echocardiography can be used to diagnose left anterior descending artery stenosis in patients with type 2 diabetes

Echocardiography, Doppler DIAGNOSES Acquired stenosis

Injury	disrupts	Physiological Function
Bodily Location	location-of	Biologic Function
Anatomical Structure	part-of	Organism
Pharmacologic Substance	causes	Pathological Function
Pharmacologic Substance	treats	Pathologic Function

Técnicas

- Patrones de captura: expresiones regulares, wrappers, reglas, árboles de decisión, etc.
- Ejemplos:
  - Clasificar por palabras en la entidad
    - Juan Carlos I→ contiene Juan que está en listado de nombres de persona luego es un persona
    - Hospital Juan Carlos I→ Aunque tiene Juan que es indicador de persona "gana" hospital que es identificador de organismo
  - Clasificar con un disparador (trigger word)
    - Sr. Juan Pérez, Prof. Madrid→ aunque Madrid pueda estar en un listado de lugares el lanzador Prof. Indica que es una persona, igual que Sr.
  - Clasificar con el contexto
    - Fuentes dice que irá → aunque Fuentes no esté en ningún listado de personas, organismos, etc. el verbo "decir" indica que probablemente sea una persona

Normalmente se resuelve con:

- **Patrones**
- **Listados de Términos**
- **Inteligencia Artificial y minería de textos**

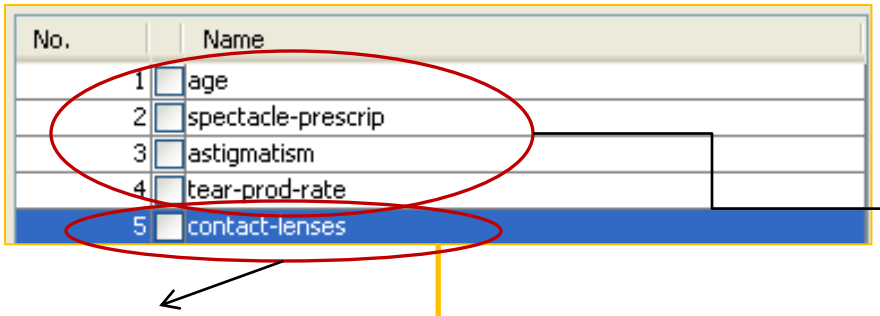
Generación manual

- Heurísticas
  - Definición de patrones a partir de observación
  - Ej. siempre que aparezca @ es correo electrónico. Patrón: "\S+@\S+"
  - Ventajas: precisión y muy ajustado a dominios específicos
  - Desventajas:
    - Mucho tiempo de elaboración
    - Requiere expertos en el área si se trata de dominios específicos
    - Son poco flexibles: difícil hacer reglas exhaustivas (baja la recall) y difícil de adaptar a nuevos contenidos

Generación automática. Aprendizaje supervisado

- A partir de observaciones se utilizan algoritmos para extraer patrones de comportamiento.
  - Elegir un corpus del que extraer las relaciones
  - Etiquetar entidades relevantes en el corpus, mejor si están en la misma sentencia
  - Etiquetar (clasificar) manualmente relaciones entre las dos entidades e indicar si no hay relación entre ellas
  - Dividir la colección en un conjunto de entrenamiento y de prueba
- Ej. determinar qué tipo de lente es adecuada para cada tipo de persona
  1. Determinar lo que quiero averiguar: tipo de lente (blanda, dura, ninguna)
  2. Seleccionar las variables que pueden influir: edad, miopía/hipermetropía, astigmatismo, producción de lágrima
  3. Recoger observaciones
  4. Generar un modelo con esos datos que dado un conjunto de variables de entrada (correspondientes a una persona) me determine qué tipo de lente es recomendable

Generación automática. Aprendizaje supervisado (II)



Age: 21
Spec-pres: myope
Astigmatism: yes
Tear-prod: reduced
Contact-len: hard
Age: …
…

```
PART decision list
------------------

tear-prod-rate = reduced: none (12.0)

astigmatism = no: soft (6.0/1.0)

spectacle-prescrip = myope: hard (3.0)

: none (3.0/1.0)

Number of Rules  :      4
```

Generación de Modelo
(con la herramienta Weka en este ejemplo)

Generación automática. Aprendizaje supervisado (III)

- En el caso de entidades
  - Se "entrena" con corpus con los tipos de entidades que nos interesa capturar anotadas.
    - Ej. si nos interesa capturar nombres de personas, el corpus vendrá con todos los nombres de personas que aparezcan anotados (el modo de anotación varía)
  - Debemos identificar en el corpus aquello que puede servir como atributo (ej. si va o no en mayúsculas, longitud de la palabra, palabra anterior, categoría gramatical, etc.)
  - La variable de clase es el tipo de entidad:
    - Ej. Identificar: entidad / no entidad
    - Ej. Clasificar: persona /organismo /localización /no entidad
  - Algunos algoritmos: Hidden Markow Models (HMM), árboles de decisión, modelos de Máxima Entropía (ME), Naive Bayes, Support Vector Machines (SVM) y condicional Random Fields (CRF)

Generación automática. Aprendizaje supervisado (IV)

- Ventajas:
  - Poco tiempo
  - Identificación de patrones difíciles de detectar
  - Buenos resultados con corpus de entrenamiento y test adecuados
- Desventajas:
  - Overfitting
  - Selección atributos
  - Modelos incomprensibles
  - Costosa la anotación de corpus de entrenamiento. Los corpus suelen ser de pago y estar sujetos a licencias. Algunos corpus anotados en: http://www.cs.technion.ac.il/~gabr/resources/data/ne_datasets.html
  - Es complicado cambiar de género

Generación automática. Otras técnicas

- Bootstrapping: se considera aprendizaje semi-supervisado, porque parte de una serie de ejemplos (ya sean patrones o entidades) y extrae de un corpus no anotado nuevas entidades.
  - Ejemplo:
    1. Se seleccionan un conjunto de entidades (semillas) de un tipo dado (ej. Bolivia, Guatemala, Honduras, con el tipo "país")
    2. Se extraen los patrones encontrados en torno a estas entidades en el corpus (ej. "oficinas en X", "instalaciones en X")
    3. Se seleccionan los mejores de estos patrones
    4. Se incorporan las mejores entidades
    5. Se ejecuta de nuevo el paso 1 añadiendo los nuevos patrones y entidades
- Uso de recursos léxicos-semánticos: listados, tesauros, etc. Uno de los más destacados: Dbpedia http://dbpedia.org

Ejemplo bootstrapping

1. Semillas "Mark Twain" y "Elmira"
2. Lanzar la búsqueda en Google (p.e.) o contra un corpus
   "Mark Twain is buried in Elmira, NY." → X is buried in Y
   "The grave of Mark Twain is in Elmira" → The grave of X is in Y
   "Elmira is Mark Twain's final resting place" → Y is X's final resting place.
3. Utilizar los patrones para buscar nuevos patrones (este paso hace que el sistemas degenere deprisa si no se controla)
4. Iterar

Se suele combinar con aprendizaje supervisado con múltiples facetas sintácticas, morfológicas, localización, tesauro, etc

Aprendizaje no supervisado

- Busca tripletas conocidas y veraces en un corpus.
- El problema es que no trabaja con una colección validada
- Es imposible calcular la exhaustividad y la precisión
- Sólo se puede extraer una muestra y calcular la precisión
- El clustering es básicamente no supervisado

Problemas genéricos en NER

- Correferencia y alias: referencia a la misma entidad de formas diferentes (ej. J. A. Abad, José A. Abad y José Antonio Abad).
    - Alias: dos entidades con nombre propio diferentes que se refiere a la misma entidad real (José Abad y Abad).
    - Referencias pronominales o anáforas (durante mi estancia)
    - "El chico del coche azul" ¿es entidad de persona?
  - Se requiere un conocimiento extra del mundo para ser resuelta.
- Identificación de la frontera o límite cuando se trata de palabras compuestas (ej. Mr. John Smith, First National Bank)
- Entre las mismas dos entidades puede haber más de un tipo de relación
- Ambigüedad con términos comunes (ej. rosa) y ambigüedad en el etiquetado (ej. "Madrid – Barcelona 1-1" ¿ciudades o equipos?)

Web Scraping

- Técnica que permite la extracción de datos de sitios web
- También llamada Web Harvesting o Web Data Extraction
- Proceso
  - Acceso a páginas web mediante HTTP o navegadores
  - Selección y copia de los datos de la página
  - Almacenamiento y procesamiento
- Ej: camelcamelcamel sobre precios en Amazon
- Utiliza el protocolo HTTP o navegadores web
- Puede realizarse manualmente o de forma automática:
  - Robots de recuperación
  - Programación de la recogida
  - Selección de datos
  - Recuperación automática de múltiples páginas

Web Scraping: técnicas

- Copiar/pegar manualmente
- Expresiones regulares (p.e. Grep)
- Protocolo HTTP para recopilación de páginas
- Análisis HTML mediante XQuery
  - Selección mediante expresiones XPATH
  - Recuperación
  - Transformación de datos
- Minería de datos/textos
- Reconocimiento de anotación semántica en páginas anotadas mediante metadatos o microformatos
  - p.e. Snippets
- Visión artificial para identificar información interpretando visualmente las páginas
- Aplicaciones específicas Web Scraping

Web Scraping: herramientas

- Aplicaciones con extracción asistida, no requieren programación
  - Import.io (https://www.import.io)
  - Octoparse: (http://www.octoparse.com
  - Screen Scraper (www.screen-scraper.com)
  - Mozenda (http://www.mozenda.com/)
  - Websundew (http://websundew.com/)
  - Web Scraper (http://webscraper.io/)
  - ParseHub (https://www.parsehub.com/)
  - Portia (https://scrapinghub.com/portia/)
  - DataScraping.co (https://www.datascraping.co/)
- Extensiones de navegadores
  - Scraper
  - Web Scraper
  - Extracty
- Frameworks para desarrollo
  - Scrapy (https://scrapy.org/) Python
  - Jsoup (https://jsoup.org/) Java

Universidad
Carlos III de Madrid

Módulo IX
Extracción de Información

Colaboradores

J.Morato, V.Palacios

J.Urbano, S.Sánchez-Cuadrado, M.Marrero