

# Extracción de Información UPF, noviembre 2004

Horacio Rodríguez

TALP Research Center  
Dep. Llenguatges i Sistemes Informàtics  
Universitat Politècnica de Catalunya  
horacio@lsi.upc.es  
<http://www.lsi.upc.es/~horacio>

## Guión de la presentación

- Elementos básicos de la Extracción de información
- Técnicas de Aprendizaje Automático en la Extracción de información
- Algunos sistemas notables
- nota
  - esta presentación se basa en un tutorial (más extenso) sobre el tema impartido por el autor en la U. Sevilla (<http://www.lsi.upc.es/~horacio/varios/sevilla2001.zip>) en junio de 2001. La presentación ha sido actualizada con material de los cursos de doctorado sobre IE de Horacio Rodríguez y Jordi Turmo

## Extracción de la Información

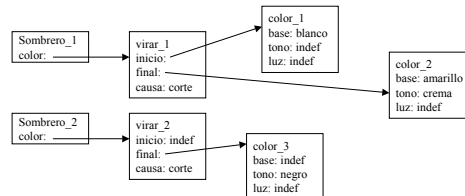
Text-Based Intelligent Systems (TBIS):  
Information {Retrieval, Integration, Mining,  
Harvesting, Filtering, Routing, Extraction,...}  
Document Classification, Question Answering,  
Topic Detection & Tracking, ...

El objetivo de la Extracción de Información consiste en localizar y extraer automáticamente piezas de información relevante para un conjunto de conceptos prescrito (el *escenario*)

## Sistemas de Extracción de Información

Típicamente un SEI extrae informaciones sobre **entidades, relaciones y eventos** a partir de documentos en un **dominio restringido**

Ejemplo: dominio micológico (M-Turbio)  
El color blanco de su sombrero pasa a amarillo crema al corte.  
El sombrero ennegrece si se corta.



## Aplicaciones

- Extracción de información de la Web
- Construcción de BD de noticias
- Integración de información
- Dominios: médico, finanzas, militar, ...
- Semantic Web
- ...

Limitaciones:  
inútil si la precisión < 90%  
alto coste de transporte y adaptación

## Ejemplo (MUC-6)

A **bomb** went off this morning near a **power tower** in **San Salvador** leaving a large part of the city without energy, but **no casualties** have been reported. According to unofficial sources, the bomb -allegedly detonated by **urban guerrilla commandos- blew up** a power tower in the northwestern part of San Salvador at 0650.

|                            |                                  |
|----------------------------|----------------------------------|
| Incident type:             | bombing                          |
| date:                      | March 19                         |
| Location:                  | El Salvador: San Salvador (city) |
| Perpetrator:               | urban guerrilla commandos        |
| Physical target:           | power tower                      |
| Human target:              | -                                |
| Effect on physical target: | destroyed                        |
| Effect on human target:    | no injury or death               |
| Instrument:                | bomb                             |

## Lecturas básicas EI

- D.E. Appelt, D.J. Israel, 1999
- M.T. Paziienza, 1997
- E. Hovy, 1999
- J. Cowie, W. Lehnert, 1996
- C. Cardie, 1997
- Y. Wilks, 1997
- J. Cowie, Y. Wilks, 2000
- R.J. Mooney, C. Cardie, 1999
- Atserias et al, 1998
- Muslea, 1999

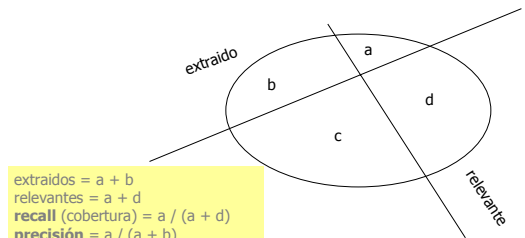
## Algunas referencias recientes

- AAAI-99 Workshop on Machine Learning for Information Extraction (1999)
  - <http://www.isi.edu/~muslea/RISE/ML4IE/>
- Artificial Intelligence Vol. 118 (2000)
  - Cohen, Craven et al, Kushmerick
- ECAI Workshop on Machine Learning for Information Extraction (2000)
  - <http://www.dcs.shef.ac.uk/~fabio/ecai-workshop.html>
- IJCAI-2001 Workshop on Adaptive Text Extraction and Mining (2001)
  - <http://www.smi.ucd.ie/ATEM2001/proceedings/toc.html>

## Historia de la EI

- Precedentes
  - Sager 1981: LSP
  - deJong 1982: FRUMP
  - Cowie 1981
  - Zarri 1983
  - Hayes 1986: JASPER
- Conferencias MUC (1987-1998)
  - Programa TIPSTER (1990-1995) (EEUU)
- Programa LRE (CE)
  - TREE, AVENTINUS, FACILE, ECRAN, SPARKLE
- Conferencias DUC (2000-)
  - Programa TIDES (1999-) (EEUU)
- PASCAL Challenge (2004-)

## Medidas de calidad de la extracción



extraídos =  $a + b$   
relevantes =  $a + d$   
**recall** (cobertura) =  $a / (a + d)$   
**precisión** =  $a / (a + b)$

**recall** = están todos los que son  
**precisión** = son todos los que están

$$F = \frac{(\beta^2 + 1) \cdot p \cdot r}{\beta^2 \cdot p + r}$$

## Conferencias MUC

- MUCK-1 (1987), MUCK-2(1989)
  - operaciones navales
- MUC-3 (1991), MUC-4 (1992)
  - terrorismo en Latinoamérica
  - estructura de salida con 18 atributos
  - cobertura y precisión
    - ok, ko, parcial
- MUC-5 (1993)
  - noticias financieras, microelectrónica
  - inglés, japonés
  - F score
- MUC-6 (1995), MUC-7 (1998)
  - Entidades propias, entidades estructuradas, correferencias, eventos

## Tipos de extracción

- Fuente textual
  - Texto libre
  - Texto estructurado
  - Texto semiestructurado
    - Páginas Web
      - marcado HTML
      - marcado XML
- Disciplinas próximas
  - Dense & Sparse Extraction
  - Topic Detection & Tracking
  - Question Answering
  - Information Integration
  - Document classification
  - Intra & Inter document linking

Wrappers

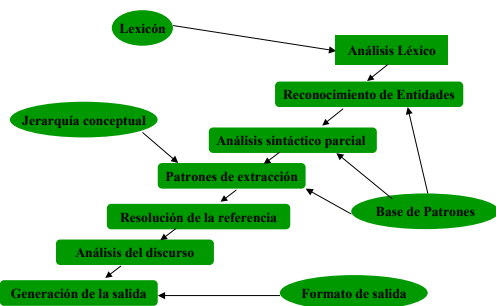
## Componentes de un SEI (Cowie, Lehnert, 1996)

- Nivel texto
  - filtrado => relevancia
- Nivel palabra
  - pos tagging
- Nivel frase
  - chunks, sintagmas, categorización semántica
- Nivel oración
  - relaciones sintácticas
- Nivel interoracional
  - coreferencia
- Nivel esquema
  - proyección sobre el formato (prescrito) de salida

## Arquitectura de un SEI

- Hobbs:
  - Cascada de transductores (o módulos) que a cada paso añaden estructura y a menudo eliminan información irrelevante aplicando reglas que son adquiridas manual o automáticamente

## Arquitectura típica de un SEI (Proteus)



## Características de los SEI (1)

- Importancia de la Ingeniería del Conocimiento
  - Modularidad
    - Tareas básicas
    - Tareas específicas
- Uso de conocimiento débil y local
- Recursos de LN
  - MRDs, Ontologías, Lexicones, Corpus, ...
- Importancia de la transportabilidad y afinado (tuning)
- Técnicas empíricas
- Salida: texto marcado o bases de datos

## Características de los SEI (2)

- Fuerte dependencia del dominio
- Estructura prescrita (Escenario)
- Semántica vs. Sintaxis
- Análisis sintáctico superficial, parcial y global
- Análisis del discurso
- Arquitectura en cascada
  - Técnicas de estados finitos
- Estructura del texto
  - metainformación
  - sublenguajes
    - género
    - dominio

## Obtención de las fuentes de conocimiento

- Conocimiento más o menos estable
  - Vocabulario general
  - Gramáticas generales
  - procesadores básicos: segmentadores, analizadores morfológicos, taggers, chunkers, parsers, ...
- Conocimiento altamente dependiente del dominio
  - Vocabulario específico del dominio, terminología
  - WSD
  - NER
  - Reglas y/o patrones de extracción

Técnicas de ML

## Análisis léxico

- (a veces) Identificación de la lengua
- Segmentación
- División del texto en unidades (tokens)
- Consulta a diccionarios
  - ej. PROTEUS (NYU)
    - Comlex, Nombres propios (personas, geográficos, empresas), ...
- Procesadores específicos
  - fechas, cantidades, siglas, locuciones, términos multipalabras, ...
- Reconocedores de nombres propios (Named Entities, NER)
  - Lexicones especializados
  - patrones (expresiones regulares)
- Palabras desconocidas

## Desambiguación morfosintáctica (pos tagging)

- Sistemas
  - basados en reglas
  - estadísticos
  - híbridos
- tagset
- calidad de la desambiguación:
  - por encima del 97%

## Análisis sintáctico

- Global
  - aproximación estándar: LaSIE, LOLITA
    - ineficiencia, limitaciones de las gramáticas, tamaño de las gramáticas
    - treebank grammars
  - aproximación en cascada: Pinocchio, Alembic
    - se solucionan algunos de los problemas anteriores
- Parcial
  - Fastus => Proteus, PLUM, PIE, Umass, HASTEN, TURBIO, ESSENCE
    - ausencia de dependencias globales.
    - uso de metarreglas para precompilar patrones

## Ejemplo PROTEUS

- Grupos nominales y verbales no recursivos (chunks)
- Grupos nominales más amplios sólo si existe evidencia semántica
- uso de metarreglas (similares a las de GPSG) para ampliar la cobertura sintáctica

## Semántica

- Normalmente sólo a nivel léxico
- A veces WSD
- Representación semántica explícita a niveles más complejos de proceso sintáctico
  - Alembic (MITRE) => forma lógica proposicional
    - M.Vilain (1999)
  - Pinocchio => quasi logical form
    - F. Ciravegna, A. Lavelli (1999)

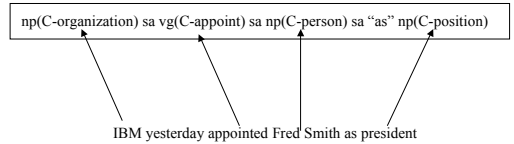
## Patrones de extracción

- las reglas de extracción suelen constar de
  - un patrón que debe aplicarse sobre la estructura (texto marcado, bosque de análisis, formas lógicas) que se ha obtenido de los procesos previos para intentar el matching
  - una o varias acciones a realizar
    - creación de ejemplares de objetos
    - relleno de atributos
    - establecimiento de relaciones

## Tipos de patrones

- 3 niveles
  - bajo nivel: gran aplicabilidad (normalmente incluidos en el sistema)
  - intermedio: librerías de patrones (aplicables a diferentes dominios)
    - ej. extractores de entidades (persona, empresa, lugar, organización)
    - extractores de relaciones (persona/oranización, organización/lugar)
  - específicos del dominio

## Ejemplo (PROTEUS)



## Adquisición de los Patrones de extracción

- Normalmente creados manualmente
- Uso limitado (pero creciente) de técnicas de ML
  - como ligar los esquemas (templates) con su expresión en LN
  - Hasta ahora Aprendizaje supervisado
  - Cada vez más Aprendizaje no supervisado
    - A. Semi-supervisado
    - Active Learning
    - Bootstrapping
    - Co-training

## Proceso discursivo

- Resolución de expresiones referenciales
  - anáforas
    - pronombres personales, posesivos, ...
  - referencias definidas
  - entidades con nombre
- Inferencias
- Integración (merging) de la información

## Ejemplos de SEI de texto libre: metodología

| Sistema  | Referencia             | Parsing                      | Semántica                            | Discurso                                   |
|----------|------------------------|------------------------------|--------------------------------------|--|
| LaSIE    | Gaizauskas et al, 1995 | Análisis en profundidad      |                                      |  |
| LOLITA   | Garigliano et al, 1998 |                              |                                      |  |
| CIRCUS   | Lehnert et al, 1991    | Chunking                     | Pattern matching                     | template merging                           |
| FASTUS   | Hobbs et al, 1993      |                              |                                      |  |
| BADGER   | Fisher et al, 1995     |                              |                                      |  |
| HASTEN   | Krupka, 1995           |                              |                                      |  |
| PROTEUS  | Grishman, 1995         |                              |                                      |  |
| ALEMBIC  | Aberdeen et al, 1993   |                              |                                      |  |
|          |                        |                              | Grammatical relations interpretation | procedimientos de interpretación semántica |
| LaSIE-II | Humphreys et al, 1998  | Partial Parsing              | lexical semantics int.               |  |
| PIE      | Lin, 1995              |                              | pattern matching                     |  |
| PLUM     | Weischedel et al, 1995 |                              |                                      |  |
| IE2      | Aone et al, 1998       | Pattern matching             |                                      | template merging                           |
| LOUELLA  | Childs et al, 1995     |                              |                                      |  |
| SIFT     | Miller et al, 1998     | parsing sintáctico-semántico |                                      |  |

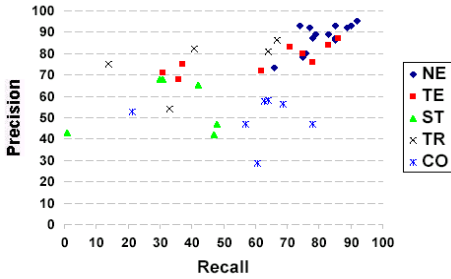
(Tomado de J.Turmo, 2002)

## Ejemplos de SEI de texto libre: Conocimiento

| Sistema  | Parsing   | Semántica                | Discurso                          |
|----------|---|--------------------------|-----------------------------------|
| LaSIE    | Treebank grammar  | $\lambda$ expresiones    |                                   |
| LOLITA   | General grammar   | semantic network         |                                   |
| CIRCUS   |   | concept nodes(AutoSlog)  |                                   |
| FASTUS   |   | hand-crafted IE rules    |                                   |
| BADGER   |   | concept nodes (CRYSTAL)  | decision tress                    |
| HASTEN   | Phrasal grammar   | E-graphs                 |                                   |
| PROTEUS  |   | Learned IE rules         |                                   |
| ALEMBIC  |   | hand-crafted G relations |                                   |
| LaSIE-II | hand-crafted stratified general grammar   | $\lambda$ expresiones    |                                   |
| PIE      | General grammar   | hand-crafted IE rules    |                                   |
| PLUM     |   |                          |                                   |
| IE2      |   | hand-crafted IE rules    | hand-crafted rules decision trees |
| LOUELLA  |   |                          |                                   |
| SIFT     | Statistical models for syntactic-semantic parsing & coreference resolution learned from PTB and on-domain annotated texts |                          |                                   |

(Tomado de J.Turmo, 2002)

## Evaluación MUC-7



## Portabilidad (1)

- Es una cualidad fundamental dada la gran dependencia del dominio de la EI
- Normalmente se deben afinar o crear de nuevo los recursos:
  - Lexicones
    - Background vs. Foreground (Kilgarrieff)
  - Ontologías
  - Base de patrones
  - estructura de salida (plantillas)

## Portabilidad (2)

- Forma de llevar a cabo el afinado
  - automáticamente
  - manualmente
  - semi-automáticamente
- La mayor dificultad (y la tarea que tiene un coste mayor) reside en la (re)construcción de la base de patrones. Por ello es aquí donde se han aplicado más esfuerzos en intentar automatizar la tarea

## Portabilidad (3)

- Afinado (tuning) de lexicones y ontologías
  - dos aproximaciones (Wilks, 1997)
    - Lexicón antiguo + corpus del (nuevo) dominio => lexicón nuevo
    - corpus del (nuevo) dominio => lexicón nuevo
  - elementos a modificar
    - palabras
    - acepciones
    - preferencias verbales (posibles alternancias de diátesis, régimen proposicional, restricciones selectivas, ...)
- Proceso
  - manual (el más corriente) con editores especializados
  - automático: E. Riloff & R. Jones (1999)

## Portabilidad (4)

- Creación o afinado de la base de patrones
  - Uso de herramientas interactivas para la adquisición manual
    - NYU Interactive tool
      - C.Nobata, S.Sekine (1998)
      - R.Yangarber, R.Grishman (1997)
    - El usuario proporciona un ejemplo (o lo extrae del corpus)
    - El usuario codifica la información a extraer a partir del ejemplo
    - El sistema utiliza la base actual de patrones para crear una descomposición estructural del ejemplo
    - Usuario y sistema interaccionan para extender y/o generalizar sintácticamente (metarreglas) y semánticamente (jerarquía conceptual) el o los patrones implicados
  - **Uso de técnicas de ML**

## Uso de técnicas de Aprendizaje Automático (ML)

- Basado en la explotación de corpus para:
  - Construcción de patrones de extracción
  - Otras tareas de bajo nivel
    - pos tagging
    - segmentación
    - chunking
    - dependencias sintácticas entre unidades
    - correferencias
  - Tareas afines o complementarias
    - resumen automático
    - clasificación de textos
- Mooney, Cardie 1999, C.Cardie (1997)

## Aprendizaje de patrones de extracción

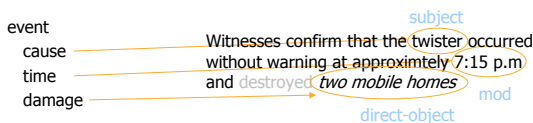
- Posibles clasificaciones de los métodos:
  - tipo de conocimiento aprendido  
(reglas, árboles de decisión, HMMs, separadores lineales, ...)
  - tipo de documentos de entrenamiento  
(texto no restringido, documentos estructurados o documentos semi-estructurados)
  - grado de supervisión  
(instance-based learning, observation-based learning, formas de active-learning, bootstrapping, ...)
  - paradigma de aprendizaje  
(propositional learning, relational learning, statistical learning, ...)
  - ...

## Aprendizaje de patrones de extracción

- Reglas
  - supervisados (la mayoría)
    - instance-based learning
      - AutoSlog en MUC-4 (Riloff, 1996)
      - CRYSTAL (Soderland et al, 1995)
      - PALKKA (Kim, Moldovan, 1995)
      - LIEP (Huffman, 1996)
      - RAPIER (Califf, Mooney, 1997), (Califf, 1998)
      - WHISK (Soderland, 1999)
      - SRV (Freitag, 1998a,b)
      - WAVE (Aseltine, 1999)
      - TIMES (Chai et al, 1999)
      - EVIUS (Turmo, Rodríguez, 2002) (Turmo, 2002)
    - requieren marcar a priori los ejemplos en el corpus de entrenamiento (proceso muy costoso)

## Aprendizaje de patrones de extracción

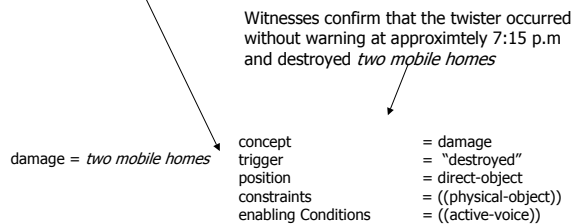
- Reglas
  - supervisados (la mayoría)
    - marcaje de ejemplos en el corpus de entrenamiento
      - generalmente preproceso de corpus  
(POS, semántica léxica y/o roles sintácticos)
      - identificar palabras activadoras
      - asociar un slot de la estructura de salida a cada elemento a extraer del ejemplo



## ejemplos (1)

### AutoSlog (Riloff, 1996)

Sistema guiado por una serie de reglas lingüísticas independientes del dominio  
extracción de **Concept Nodes**



## ejemplos (2)

### proceso:

- 1) generar el corpus de aprendizaje apropiado (información etiquetada con etiquetas semánticas)
- 2) identificar roles sintácticos de las partes etiquetadas
- 3) identificar las palabras activadoras (trigger words)
- 4) proceso de aprendizaje guiado por un paquete de heurísticas que actúan sobre las palabras activadoras y su contexto inmediato

### AutoSlog-TS

prescinde de la supervisión. La intervención humana se limita a clasificar de relevante o irrelevante el texto que se incorpora al proceso de aprendizaje

## ejemplos (3)

### CRYSTAL (Soderland et al, 1995)

Utiliza técnicas de formación de conceptos (Concept Induction Learning Michalski).

Dominio médico (utiliza la jerarquía semántica de UMLS).

Usa corpus anotado para el aprendizaje (analizado sintácticamente).

Generaliza a partir de contextos lingüísticos especificados con gran detalle.

Aproximación ascendente. Se relajan gradualmente las restricciones sobre la definición inicial (máxima especificidad) de forma que se amplía la cobertura incorporando los conceptos más similares (integrando sus definiciones) para lograr un diccionario más compacto

## ejemplos (4)

### WHISK (Soderland, 1999)

Aprendizaje de patrones expresados como expresiones regulares de forma que es posible la extracción simultánea de varios descriptores

Inducción de reglas

Inducción descendente iniciada por un ejemplo específico  
Uso de clases semánticas dependientes del dominio para clasificar las palabras

Aplicación a:  
texto libre  
texto marcado HTML  
texto previamente analizado sintácticamente

## ejemplos (5)

### RAPIER (Califf, Mooney, 1997), Califf (1998)

Robust Automated Production of Information Extraction Rules  
Aprendizaje de patrones expresados como expresiones regulares

pre-filler pattern  
filler pattern  
post-filler pattern

Algoritmo ILP que actúa sobre el texto asignado a cada descriptor y su contexto (ilimitado) inmediato.

Utiliza el texto con etiquetado morfosintáctico desambiguado (tagger de Brill)

Utiliza WordNet

## ejemplos (6)

### ejemplo de RAPIER

"... sold to the bank for an *undisclosed* amount..."

"... paid Honeywell an *undisclosed* price..."

| Pre-filler                                | Filler                      | Post-filler   |
|---|-----------------------------|---------------|
| 1) POS: {nn, nnp}<br>2) List: maxlength 2 | 1) "undisclosed"<br>POS: jj | 1) Sem: price |

## ejemplos (7)

### SRV (Freitag, 1998ab)

Aprendizaje relacional (derivado de FOIL)

Rasgos simples (atributos) y relacionales

Relaciones sintácticas: Link Grammar  
Relaciones semánticas: WordNet

Aplicado (entre otros) a la clasificación de páginas Web

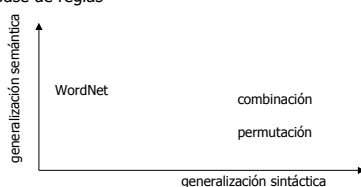
En Freitag, 1998b se combinan tres estrategias de aprendizaje

## ejemplos (8)

### TIMES (Chai et al, 1999)

Trainable InforMation Extraction System.

Algoritmo de fuerza bruta. A partir de cada ejemplo proporcionado por el usuario el sistema propone en forma automática un serie de posibles generalizaciones. Cuando una regla propuesta supera una cota de cobertura en el conjunto de entrenamiento, el sistema la incorpora a su base de reglas



## ejemplos (9)

### (Riloff, Jones, 1999)

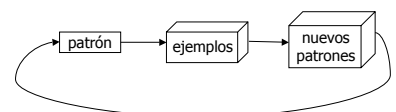
Mutual Bootstrapping.

Aprendizaje simultáneo de un lexicón semántico (dominio) y de la base de patrones (escenario).

Utiliza un corpus no anotado.

Para cada clase semántica se define (manualmente) un conjunto inicial de palabras (*seed words*).

Ampliación: Multi-Level Bootstrapping.



## Aprendizaje de patrones de extracción

- Reglas

menos supervisados (probados para texto no restringido)

- observation-based learning
  - AutoSlog-TS (Riloff, 1996)
    - observación= texto relevante/irrelevante
    - basado en AutoSlog
  - ESSENCE (Català, 2000)
    - observación= ventana de k palabras centrada en una palabra clave (nombre o verbo)
    - generalización de observaciones mediante un algoritmo de cobertura bottom-up

## Aprendizaje de patrones de extracción

- Reglas

menos supervisados (probados para texto no restringido)

- active learning: ExDISCO (Yangarber, 2000)
  - conjunto inicial de reglas construidas a mano
  - selección de los documentos de entrenamiento que contienen extracciones
  - puntuación de las extracciones
  - selección de la mejor extracción como nuevo ejemplo para el aprendizaje incremental
- Counter-Training: (Yangarber, 2003)

## Conclusiones

- Línea de investigación sumamente activa
- Técnicas de tratamiento sumamente variadas
  - simbólicas, estadísticas, ML, ...
- Uso creciente de técnicas de ML
  - PASCAL challenge 2004
- Comparte técnicas y recursos con otras áreas de PLN
- Aplicaciones actuales y futuras numerosas en el campo del PLN