Índice

Extracción de Información (EI)

- A.k.a. *Information Extraction (IE)*
- **Def.:** área de la ciencia y la tecnología que trata de la identificación, clasificación y estructuración en clases semánticas de información específica encontrada en fuentes desestructuradas (p.ej., textos), para así permitir su posterior procesamiento automático en tareas de procesamiento de la información.
- **Objetivo:** dada una **colección de documentos**, identificar y extraer de los mismos aquellos hechos relevantes para un dominio particular (*dominio de extracción*), ignorando la información extraña e irrelevante.
  - La información obtenida se devuelve de forma **estructurada**.
  - Ha de establecerse a priori qué constituye un hecho relevante.
  - Sistemas especializados de **dominio acotado**.

Extracción de Información (EI): Ejemplos

- Una compañía quiere realizar un seguimiento de las reacciones a su nuevo producto en diferentes *blogs*
- Una compañía abonada a un proveedor de noticias económicas quiere realizar un seguimiento de las reacciones del mercado a los rumores sobre quiebras, fusiones y opas de empresas en bolsa. Dicha información será organizada cronológicamente y por compañía.
- Una agencia de seguridad desea hacer un seguimiento del tráfico de *mail* en busca de indicios de actividades delictivas.
- Un equipo de investigación farmacéutica quiere analizar toda la literatura disponible para conocer todas las interacciones de un cierto grupo de proteínas con otras proteínas.
- *Shopbots*: buscan y comparan precios de productos en diferentes webs

Datos Estructurados vs. Desestructurados

- **Estructurados**: aquéllos de semántica definida y susceptibles de ser procesados automáticamente por el ordenador (p.ej., bases de datos)
  - Bases de datos, hojas de cálculo, etc.
- **Desestructurados**: aquéllos donde la información está codificada de forma que no permite su procesamiento automático inmediato
  - Texto escrito o hablado, grabaciones radiofónicas, etc.
- **Objetivo:** obtener información estructurada a partir de textos en lenguaje natural (i.e. desestructurados)

"Vendo Peugot 205 con 100.000 km. 6500 euros. Tlf 981123456. Llamar después 20:00."

Modelo	Precio	Tel
Peugot 205	6500	981123456

Extracción de Información Semántica

- **Ppo. de composicionalidad de Frege:** "la representación semántica de un objeto puede obtenerse a partir de las representaciones semánticas de sus componentes".
- **Cadena de realización** *(realizational chain)*: en un lenguaje dado la estructura superficial (texto) es fruto de sucesivas etapas de transformación a lo largo de diferentes niveles de abstracción partiendo de su significado último y original:

  idea $\dashrightarrow$ conceptos semánticos de sus componentes $\dashrightarrow$ conceptos gramaticales y léxicos $\dashrightarrow$ texto

- PLN (EI) considera que este proceso es **bidireccional**: podemos aproximar la semántica de un texto a partir de sus regularidades a nivel superficial
  - i.e., aplicaremos **patrones** sobre el texto para identificar y extraer la información relevante

Especificidad de la Información

A tres niveles:

1. **Tipo de información (semántica) a extraer**: fijada a priori
   p.ej. bancarrotas y cuándo se producen
   - Las formas de expresar un evento/información son limitadas
     p.ej. concepto de 'quiebra' + expresiones temporales
   - Consecuentemente, se puede diseñar un método para identificarlos
2. **Unidad de extracción**: no se devuelve el documento completo, sino frases simples (gen. frases nominales) u otras unidades de texto a especificar
3. **Alcance de la extracción**: debe especificarse si la información puede ser extraída o no de diferentes cláusulas, oraciones, párrafos o textos

```
*Martinsa Fadesa* sucumbe a la crisis
Publicado el *15-07-2008* , por Fulanito
El consejo de administración de *la inmobiliaria* acordó *anoche* por
unanimidad presentar *concurso voluntario de acreedores* ante la (...)
```

Clasificación y Estructuración (I): Clasificación

Objetivo: convertir la información desestructurada inicial en información estructurada lista para ser procesada

- **Clasificación**
  - Una vez extraída, la información es clasificada (semánticamente)
  - Objetivo: información semánticamente bien definida
  - Condición: necesario *esquema de clasificación* (i.e. un conjunto de clases organizadas y bien definidas; p.ej. jerarquía)

personas
lugares
compañías
cargos
organizaciones
(...)

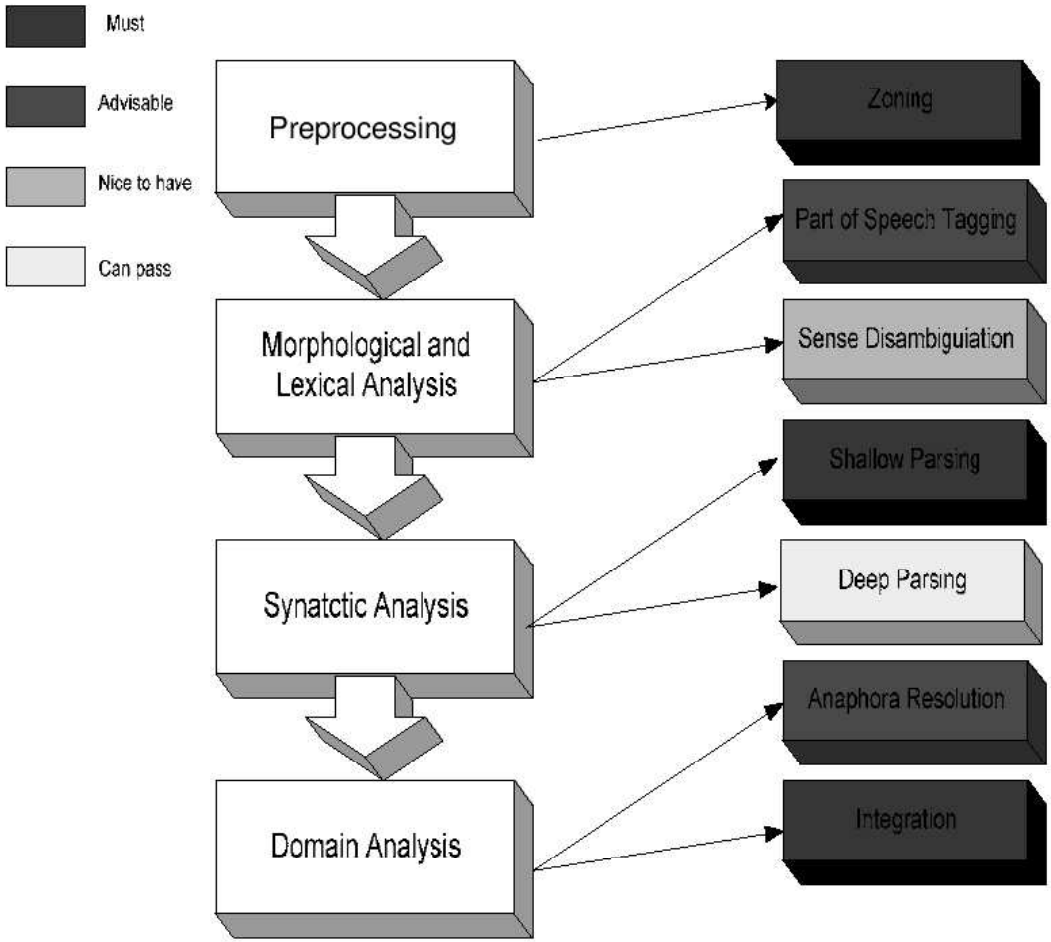
Clasificación y Estructuración (II): Estructuración

- **Estructuración**
  - La información obtenida debe almacenarse de forma estructurada
  - Solución: **plantillas (templates)**, estructuras tipo *frame* formadas por pares atributo-valor (*slots*) correspondientes a aspectos relevantes de ese evento
  - Objetivo: ir rellenando la plantilla mapeando en los diferentes slots la información contenida en el texto procesado

```
Bridgestone Sports Co. said Friday it has set up a joint venture
in Taiwan with a local concern and a Japanese trading house
to produce golf clubs to be shipped to Japan.
The joint venture, Bridgestone Sports Taiwan Co., capitalized at
20 million new Taiwan dollars, will start production in January 1990.
```

Relationship:	**TIE-UP**
Entities:	**"Bridgestone Sports Co."**
	"a local concern"
	"a Japanese trading house"
JV Company:	**"Bridgestone Sports Taiwan Co."**
Capitalization:	**20000000 TWD**

Arquitectura Genérica



Cascada de módulos que en cada paso ...

- agregan estructura al documento
- filtran información relevante por medio de la aplicación de reglas

Preprocesamiento

- **Delimitador** (*text zoner*): dividir un texto en segmentos de texto, por ejemplo párrafos.
- **Segmentador–tokenizador**: dividir los segmentos en oraciones y palabras
- **Filtro** (*filter*): elimina las oraciones no relevantes

Procesamiento Morfológico y Léxico

- **Etiquetación** (*Part-of-Speech tagging*): obtención de la etiqueta morfosintáctica de una palabra
- **Lematización**: obtención del lema (forma canónica) de una palabra
  - *Stemming* como alternativa
- **Desambiguación del sentido de la palabra** (*Word Sense Disambiguation*): en el caso de palabras polisémicas, identificar el significado/sentido concreto en ese contexto
- **Detección y análisis de entidades** (*entity recognition*): nombres propios, fechas, cifras, etc.

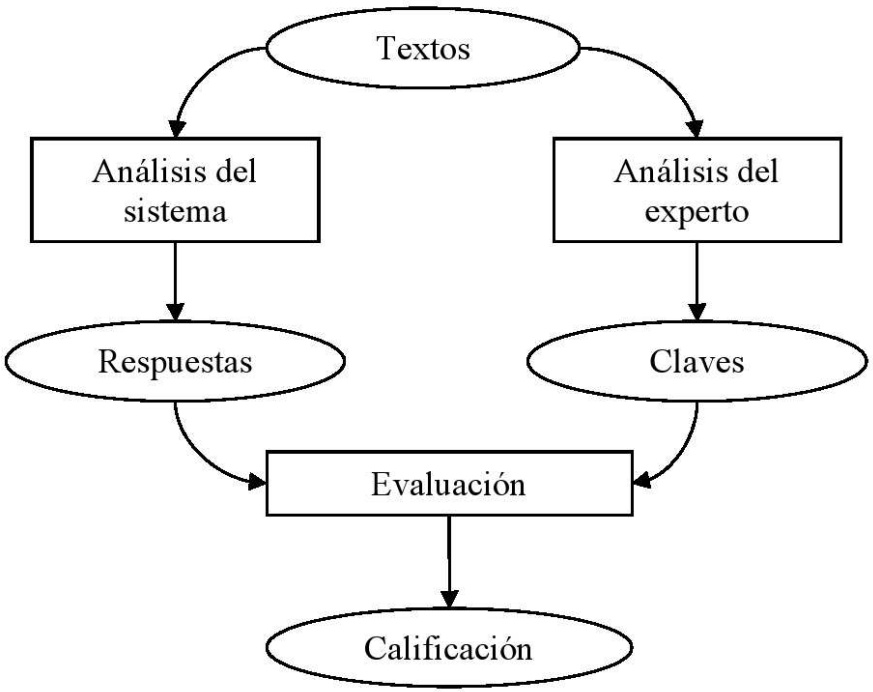
Análisis Sintáctico

- **Análisis sintáctico completo/clásico** (*full parsing*)
  - Técnicas dinámicas (p.ej. algoritmo de Earley)
  - Problemas:
    - Requiere conocimiento/recursos lingüísticos complejos (gramáticas, treebanks)
    - Escasa cobertura de las gramáticas
    - Escasa robustez
    - Alto coste
- **Análisis sintáctico superficial** (*shallow parsing* o *chunking*)
  - Obtención de constituyentes/unidades gramaticales de la oración (frases nominales, verbales, preposicionales, ...) sin detallar la estructura arborescente
  - Requerimientos menores
  - Mayor robustez
  - Bajo coste

Análisis del Dominio

- **Resolución de co-referencias**: identificar y resolver las expresiones del texto que hacen referencia al mismo objeto: anáforas, pronombres, expresiones temporales relaticas ("hoy", "hace dos días", ...), etc.
- **Tratamiento de la elipsis** (i.e. omitir en una oración una o más palabras)
- **Combinación de resultados parciales**:
  - Diferentes oraciones/documentos pueden referenciar el mismo suceso
  - Combinar la información obtenida de cada uno
- **Generación de plantillas de salida**: enlazar los elementos de información extraídos con el formato de salida especificado
  - Umbral mínimo de "interés" del evento: desechamos eventos que no lo cumplen
  - Puede ser necesario adaptar el elemento extraído al registro destino: p.ej., registros con conjunto de valores predefinido, normalización de fechas/cantidades, etc.

Proceso de Evaluación



3 elementos:

1. **Textos**: colección de docs. de los que extraer la información
2. **Claves (keys)**: conjunto de registros extraídos por los expertos (i.e. de referencia)
3. **Respuestas (responses)**: conjunto de registros extraídos por el sistema (i.e. a evaluar)

Métricas de Evaluación: Casuística

correcta	respuesta $=$ clave
parcial	respuesta $\cong$ clave
incorrecta	respuesta $\neq$ clave
espúrea	SÍ respuesta, NO clave
perdida	NO respuesta, SÍ clave
evasiva	NO respuesta, NO clave

Métricas de Evaluación: Casos de Error

- **Error en respuestas (error per response fill)**: error total

$$error = \frac{\#incorrectas + \#parciales/2 + \#espureas + \#perdidas}{\#claves + \#espureas}$$

- **Subgeneración (undergeneration)**: porcentaje de registros sin extraer

$$undergeneration = \frac{\#perdidas}{\#claves}$$

- **Sobregeneración (overgeneration)**: porcentaje de respuestas "de más"

$$overgeneration = \frac{\#espureas}{\#respuestas}$$

- **Sustitución (substitution)**: porcentaje de respuestas "cambiadas"

$$overgeneration = \frac{\#incorrectas + \#parciales/2}{\#correctas + \#parciales + \#incorrectas}$$

Métricas de Evaluación: "Clásicas"

- **Precisión (precision)**: porcentaje de respuestas correctas

$$precision \;\; Pr = \frac{\#correctas + \#parciales/2}{\#respuestas}$$

Capacidad para extraer sólo registros correctos.

- **Cobertura (recall)**: porcentaje de registros extraídos

$$recall \;\; Re = \frac{\#correctas + \#parciales/2}{\#claves}$$

Capacidad para extraer todas los registros correctos.

- **Medida-F o F_1 (F-measure/balanced F-score)**: combina ambas

$$F_1 = \frac{2\ Re\ Pr}{Re + Pr}$$

Valora Re/Pr por igual.

MUC

- Message Understanding Conference
  (`http://www-nlpir.nist.gov/related_projects/muc/index.html`)
  - Defense Advanced Research Projects Agency (DARPA)
- **Objetivo:**
  - Promover el I+D en tareas de IE
  - Facilitar infraestructura, herramientas y metodologías para la **evaluación de sistemas de IE**
- **Dominio de trabajo**:
  - MUC-1..2: comunicaciones militares navales
  - MUC-3..4: noticias de ataques terroristas
  - MUC-5: noticias de fusiones de empresas; anuncios de productos de microelectrónica
  - MUC-6: noticias de sucesiones en la dirección de empresas
  - MUC-7: noticias de accidentes de avión; noticias de lanzamientos de cohetes y misiles

MUC (cont.): Tareas de evaluación

- **Generación de plantillas de escenario**: i.e. salida del proceso completo (tarea original)
- **Reconocimiento de entidades (entity recognition)***: encontrar y clasificar las entidades, e.g., nombres de personas, organizaciones, lugares, expresiones temporales y numéricas.
  - Ampliación a multilingüe (Multilingual Entity Task, MET): español, japonés y chino (`http://www-nlpir.nist.gov/related_projects/tipster/met.htm`)
- **Resolución de correferencias***: identificar las expresiones en el texto que hacen referencia al mismo objeto
- **Plantillas de elementos***: añadir información descriptiva al resultado del reconocimiento de entidades, i.e., estandarizar conceptos (e.g. persona y organización)
- **Relación de plantillas***: identificar las relaciones entre las plantillas de los diferentes elementos, e.g., 'empleado de', 'localizado en' y 'producto de'

(*) sólo en las últimas ediciones